Going Mobile: Technology and Policy Issues in the Mobile Internet

Richard Bennett

Information Technology and Innovation Foundation
Washington, DC

March 2, 2010

Executive Summary

Ten years ago, the typical network experience was limited to dialing-up a walled garden to see if we had mail and then poking around a few familiar parts of the Internet. The advent of broadband networking changed this dramatically: it sped up the Web and brought about a host of innovative new applications and services. While the Internet was permeating modern life, a parallel development was taking place that would have even greater significance for billions of people around the world: the development of the cell phone. Inevitably, these two transformative technologies began to merge, enabling the rise of the Mobile Internet.

Convergence is now beginning to rebuild the Web into a personalized, real-time system that responds to the locations, tastes, and whims of billions of people as they live their lives at their desks, in their living rooms, or moving through the world with the flow of experience. Over the next decade, many more people will use the Mobile Internet, and it will produce an enormous array of innovations and quality of life benefits.

Even with all the network magic we enjoy today, we're still at a very early stage in the development of the Mobile Internet; with any luck, we'll look back in another ten years and wonder how we could ever have been so naïve as to tolerate the limitations of the network experience we enjoy today. The flowering of the Mobile Internet will only come to pass, however, when engineering and policy collaborate to successfully overcome the challenges to the development of a Mobile Internet that lives up to its full potential. For this to happen, policymakers must do two key things: First, they need to refrain from strangling the Mobile Internet with excessive regulation, realizing that the well of

innovation that brought us where we are has not run dry. Second, policy makers need to ensure that the mobile Internet can develop the infrastructure it needs, the most important part of which is spectrum. Policymakers need to make tough choices, transferring spectrum from less compelling historical uses to the emerging Mobile Internet.

This report examines changes that must be made to the Internet and to the mobile network to make the Mobile Internet a pervasive and universal reality in the United States and the rest of the world. Some of these changes are purely technical, but their scope affects Internet engineering as well as mobile network, device, and application engineering. The rest of the changes will take place in the policy sphere, affecting notions of network neutrality and spectrum policy. The examination of technology is quite extensive, and is illustrated with specific examples of emerging devices and applications.

In order to make effective policy for the mobile Internet it's necessary to understand the development of the Internet, the dynamics of the mobile network, and how the converged Mobile internet differs from both of these ancestors. While the traditional Internet and the Mobile Internet share some common features, they operate very differently. The traditional Internet was designed for a small group of low duty cycle, occasional use applications for locked-down computers shared by a modest number of highly-skilled, trustworthy users in a non-commercial setting; but mobile telephony's heritage is one of real-time communication-oriented applications, a diverse group of mobile users, and personal devices competing for market share. It's not surprising that there's friction between Internet culture and mobile culture.

The mobile network was originally designed to serve as an extension of the telephone network that added mobility at the network edge without altering the telephone network's fundamentals. Initially, it used analog technology, and converted to digital in the 1990s. Data services were a special feature added on to the mobile network roughly during the period of its transition from analog to digital. As presently operated, the mobile network is still more efficient at providing telephone service than data service.

Mobile data rates have doubled roughly every 30 months, as predicted by Cooper's Law. By way of contrast, Butter's Law predicts that the data rate of optical fiber doubles every nine months. Because of this, some have said that wireless is a generation behind wired systems and always will be.

This is important because the rate of progress for Internet applications is largely driven by price/capacity improvements in physical networking since the Internet protocols have been stagnant since 1993. As Internet use shifts to wireless networks with slower intrinsic rates of advance, we might expect a slower overall rate of innovation. We might also expect increased fracture between mobile applications and stationary ones, in part because the bandwidth gap between the two can only grow larger. Nevertheless, the benefits of mobility are so great that the rate of Mobile Internet innovation is bound to increase beyond anything we've seen so far, bandwidth constraints notwithstanding.

Mobile networks require more extensive management and tuning than wired networks, as their capacity is relatively more constrained and demand for this limited capacity is more variable because of roaming. Mobile networks differentiate packets by application, providing very different routing and processing to voice packets than to data packets. This differentiated treatment is a reflection of application requirements; the need for it will persist after mobile networks are fully integrated with the Internet.

While the design and engineering challenges to the full integration of the Internet with the mobile network are serious, considerable progress has been made and the path to success is reasonably clear. The Mobile Internet is already emerging, and with it an exciting new category of applications known as Mobile Augmented Reality.

Operational challenges to the adoption of the Mobile Internet are also well understood, but less easily solved. Networks operators need to build more base stations, add more radio sectors to existing base stations, and increase backhaul bandwidth. While these challenges are relatively simple in the suburbs and exurbs – all it takes is money and accommodating local governments – they're much more difficult in central cities and in rural areas. Next generation systems such as LTE consume more bandwidth than traditional cellular, which requires a beefed up middle mile. Increased use of fiber to connect cell towers with operator facilities and on to Internet exchanges may have positive spillover effects for residential broadband as more dark fiber is deployed.

There are two key policy issues for the Mobile Internet: net neutrality and spectrum. The net neutrality proceeding currently before the FCC – the Open Internet Notice of Proposed Rulemaking – proposes to envelope the Mobile Internet within the same, highly stringent, regulatory umbrella as the wired Internet. Harmonized regulation is philosophically appealing, but has a number of practical drawbacks. If the framework itself were clear and fundamentally sound, a common regime would make sense: after all, the Internet is not as much as wired network as a virtual network and its structure is meant to be technology neutral. However, if the approach is based on preserving wired network operational norms, as it currently is, then the common umbrella becomes a common straightjacket, undesirable for both wired and mobile networks.

Spectrum policy has historically featured conflict between licensing regimes and unlicensed "Open Spectrum" models such as the White Spaces system. With the parties to this controversy in détente, the focus shifts to the struggle among various license holders. The United States unfortunately adopted an obsolete standard for Digital TV ten years ago, and has failed to reap as large a digital dividend as Europe and Asia will gain as they transition away from analog television. Extracting poorly utilized DTV spectrum from broadcasters is a daunting challenge that must be solved by federal regulators with all the creativity they can muster. It's unfortunate that TV broadcasting casts such a long shadow on mobile networking at a time when 85% of Americans watch TV over a cable or satellite system and few of the over-the-air subscribers watch on HD screens. The broadcast filibuster can be mitigated by offering incentives for broadcasters to share spectrum with each other and give back the excess for auction, and by modernizing government's spectrum use.

The general approach we recommend is for the government to facilitate the Mobile Internet by removing impediments to further build-out and adoption. Speculative fears have played too large a role in the Internet regulation debates of the last decade, and it's more productive to shift the emphasis toward the government's role in facilitating progress.

First, it would be a mistake to impose the "net neutrality heavy" guidelines on either wired ISP networks or mobile networks. Rather than enacting overly prescriptive regulations banning experiments with new transport services and business models, the FCC should rely primarily on transparency and disclosure to protect consumers from speculative harms, maintain active oversight of provider practices, and reserve direct intervention for instances of clearly harmful conduct. Second, policymakers should embark on a program of spectrum modernization and expansion to ensure that mobile services can continue to grow. A special focus should be placed on the transfer of licenses from inefficient DTV use to the general pool of spectrum available for auction.

Spectrum modernization should also be employed to replace inefficient federal, state and local government uses and release unneeded spectrum to an auction pool. Finally, regulations should encourage technical solutions to be developed and deployed that enable consumers to obtain the best possible service for the best prices. Doctrinaire net neutrality heavy formulas simply don't accomplish these ends for mobile networks.

1. Stick with Light-touch Regulation

If heavy-handed net neutrality regulation is ultimately bad for investment, deployment, and adoption of wireline networks, as it is, it is potentially a fatal disaster for mobile networks. A key way to ensure that networks serve the public interest is through market mechanisms based on meaningful competition. The United States enjoys among the most competitive intermodal wireline broadband and even stronger wireless competition, with four national wireless networks, as well as a number of regional networks and Mobile Virtual Network Operators (MVNOs) such as Virgin Mobile. Fixed wireless networks such as Clearwire and the emerging LTE system are both reasonable substitutes for wired broadband, and the two satellite networks are in the process of upgrading capacity significantly. Competition can be made more effective by ensuring there are minimal delays in switching between mobile providers.

2. Enact a Sensible Transparency Rule

Just as a well-functioning democracy requires an informed citizenry, a well-functioning network ecosystem requires its well-informed and honest critics. While the new European Internet transparency rule is too new to be judged a complete success, it represents a promising direction for which there is broad consensus. There is still disagreement regarding the specific nature of required disclosure, which is understandable given the complexity of network systems and the gap between consumer awareness and technology. The challenge for a transparency rule is to disclose the things that must be disclosed in order for users to gauge the experience they'll have on any given part of the Internet ecosystem in terms the average person can understand, while making additional information available to the audience of technologists and policy analysts. Certain details of practice represent trade secrets and need not be disclosed; the means by which a particular user-visible effect is produced are less important than the effect itself. One approach that recommends itself is the co-regulatory approach championed by Marsden, in which stakeholders convene with the regulator to draft specific guidelines.[1] Toward that end, we encourage stakeholders to form a working group to advise the FCC on the particulars of disclosure.

3. Define Reasonable Network Management

The transparency rule, and its specific implementation, provides insight into the boundaries of reasonable network management practices. While the use of the term "reasonable" without definition is impossibly vague, anchoring management practices to service disclosure resolves a great deal of the mystery. We know that a practice is reasonable if it does what the operator says it does, conforms to standards devised by responsible bodies such as IEEE 802, IETF, and the ITU, and doesn't violate basic user freedoms. We know that it's unreasonable if it fails to accomplish its stated purposes, arbitrarily restricts the use of applications, or restricts basic user rights. Beyond these general guidelines, a Technical Advisory Group must work with the FCC to develop additional clarity regarding management boundaries and help advise on a case-by-case basis when needed.

4. Legitimize Enhanced Transport Services

There is widespread agreement among filers in the FCC's Open Internet NPRM that differentiated services for differentiated fees are legitimate in their own right, and not simply as an adjunct to network management. Similar services have a long history on the Internet, where they are known as Content Delivery Networks, Overlay Networks, and Transit Networks. The logic of "pay more to get more" has long been accepted practice. These practices have proved worthwhile for content resellers and application service providers such as Netflix and Skype, so it stands to reason that they would be beneficial for future competitors in the market for video streaming and telephony. If ISPs who operate the so-called "eyeball networks," including wireless mobile Internet services, serving retail customers are permitted to compete with CDNs and Overlays, new application entrants can expect lower prices and more competition, and end users can expect a wider array of options, especially among mobile applications.

5. Preserve Engineering and Operations Freedom

The primary emphasis of the Open Internet NPRM's framework of rules is on the preservation of users' freedom to experience the Internet as they see fit, without arbitrary limitations. A key way to preserve this freedom is to address the dynamics of technical freedom that make it possible. Users experience the Internet as they do now because engineers, network operators, and application innovators have been free to improve networks, network technology, and user experience.

Toward that end, the NPRM should make it clear nothing in the FCC's approach denies the freedom to invent, develop, and adopt new networking technologies, business models, and practices that have the potential to enhance the Internet's power, efficiency, vitality, or effectiveness.

To operationalize this, the FCC should consider adding two additional principles to its list: Engineering Freedom and Operations Freedom. The telephones that worked on the PSTN in the first year of the Carterfone regime still work 35 years later. If the cell phones we use today are still usable on the mobile network 35 years from now (or even ten years from now), that should be regarded as a failure of innovation. The Mobile Internet is driven by an ethic of continual improvement and this principle more than any other must remain in the forefront. Thus, we propose two additional rules for the Open Internet NPRM:

- No part of this regulation shall be construed as limiting the freedom of network engineering to devise, develop, and deploy technologies to enhance the Internet or to improve user experience.

- No part of this regulation shall be construed as limiting the freedom of Internet Service Providers, other network operators, or other service providers to devise new financial or business models that better align user incentives with those of network operators or application-based service providers without limiting user choice.

These rules make it clear that innovation is the engine that best ensures the Internet's continued public value.

6. Review Existing Spectrum Licenses

The FCC needs to complete its inventory of the licenses it has issued over the years, and implement a system that eliminates or reduces ambiguity about licenses going forward. If it's true that the FCC has somehow lost track of some licenses, as some have suggested, this error should be corrected. It's simply not acceptable for the national regulator of wireless networks to lose track of issued licenses. Legislation to create a national spectrum map introduced by Sen. Kerry (D-MA) and Sen. Snowe (R-ME), is a step in the right direction.

7. Eliminate Redundant and Archaic Licenses

Once the license inventory is complete, it will be possible to examine licenses to determine which are unused, which are redundant, and which can be combined with others to free up spectrum for auction or other kinds of assignment. Part of this process will entail reassigning some occasional uses to the control of other agencies, license holders, or custodians of other kinds. Rarely used public safety applications can be combined with consumer services, for example, by allowing public safety uses to take precedence in times of emergency. The general principle that should hold in the process of review is modernization, replacing archaic analog applications with more spectrum-efficient digital ones. No single approach to spectrum management exceeds all others in terms of general utility, but there should be a bias in favor of spectrum custodians in either the public or the private sector with vested interests in efficient use. Sufficient spectrum exists, in principle, to meet projected user requirements for mobile networking. There is not sufficient spectrum that we can afford to waste large swathes on speculative projects of uncertain utility, however. A reasonable approach is embodied in the White Spaces order, where all licenses are experimental ones renewable day-by-day. Proven applications can be rewarded under this system with license of longer duration.

> Spectrum grants for digital TV greatly exceed consumer demand and should be reduced in the public interest.

In addition, spectrum grants for DTV greatly exceed consumer demand and should be reduced in the public interest with the freed up spectrum auctioned off. Spectrum policy should respect the public's evident wishes and make more spectrum available for Mobile Internet services for which demand is growing.

8. Protect Spectrum Subleasing

Secondary markets for licensed spectrum enabled by resale and subleasing have proved useful in the U. S., where dozens of Mobile Virtual Network Operators (MVNOs) lease capacity from license holders and roaming agreements permit licensees to share capacity. These kinds of secondary markets are also useful in the microwave backhaul and point-to-point space where a given license holder can adjust microwave paths with relays and dogleg arrangements to accommodate most effective use. Therefore it is important for policy to permit the trading and leasing of most licensed spectrum.

9. Cautiously Enable Secondary Uses

One area of controversy concerns such secondary uses as wireless underlay and overlays on licensed spectrum. Advocates insist that such uses are non-interfering with properly restricted, and license holders are skeptical. The reality is that the nature of the interference caused by overlay networks such as Ultra-Wideband depends on the nature of the incumbent service. Ultra-Wideband interferes, in some installations, with highly sensitive applications such as radio astronomy, but this fact is known and the Ultra-Wideband waveform is adjusted accordingly. When the details of the incumbent service are known, in terms of coding, modulation, and framing protocols, overlay and underlay services can be engineered for cooperation without interference. Nevertheless, when details of the primary service change, interference may arise anew. For this reason, all secondary uses should be required to back off and even shut down completely until they can be certified as non-interfering with the primary license holder. The principle use of secondary services should be in areas where the primary user is not active; this is the logic behind the Dynamic Frequency Selection (DFS) system in IEEE 802.11a Wi-Fi. This system requires Wi-Fi systems to look for the use of radar on certain channels, and to refrain from using channels where radar is found.

In all cases, the burden falls on the secondary user to avoid causing interference with the primary user. Systems of enforcement for this principle need to be incorporated into all secondary use regulations; the White Spaces database has this capability.

10. Allow the Experiment to Continue

The Internet as we know it today is the fruit of a 35-year experiment. In the beginning, it was the prototypical science project, albeit one with government support shepherded by a highly skilled and dedicated band of researchers, champions, and developers out to prove that a new vision of networking was not only practical but superior to the old one.

The mobile data network has a completely different creation story, originating in a commercial context and targeted toward adding an important new feature to the existing network without fundamentally altering its nature.

Each of these networks has a story, a set of champions, and a vision. Each has been transformative in its own way, giving rise to its own industry, and liberating some vital element of human society along the way. It's not surprising that the convergence of these networks should occasion debate and conflict, some of it intense and heated.

The way forward requires some give and take. It's not enough to impose the Internet's operational traditions on the mobile network, because the Internet's operational community has chosen not to adopt the Internet standards most relevant to mobile networking: RSVP, IntServ, and Mobile IP. It's not enough for mobile operators to demand that Internet users abandon open access to the web at reasonable speeds in favor of a constrained system of locked-down portals and proxies. Each culture has things to learn from the other.

The way forward is a careful, diligent, step-by-step process beginning with reviews of historical rules and precedents and ending in the creation of a new framework that will enable the next generation of networking to flourish. The evidence of an emerging consensus among responsible parties in the United States and Europe suggests it's well underway.

Table of Contents

EXECUTIVE SUMMARY 1

1. INTRODUCTION 13

1.1. Advent of Broadband 13

1.2. Rise of the Cell Phone 14

1.3. Convergence 14

1.4. Reader's Guide to the Report 15

2. ISSUES IN INTERNET DESIGN AND OPERATION 15

2.1. Brief History 15

2.2. Modularity 16

2.3. Efficiency 18

2.4. Manageability 19

2.5. Innovation 20

2.6. Subsidies 22

2.7. Specific Example: Inter-Domain Quality of Service 24

2.8. Why the Internet Works 26

3. ISSUES IN MOBILE NETWORK DESIGN AND OPERATION 28

3.1. Brief History 28

3.2. Modularity 30

3.3. Efficiency 31

3.4. Manageability 32

3.5. Innovation 35

3.6. Specific Example: Voice over LTE 38

3.7. Status of Internet Design and Wireless Networks 40

4. THE MOBILE NETWORK INFRASTRUCTURE 41

4.1. Essential Parts of the Mobile Network 41

4.2. Spectrum 42

4.2.1. Licenses 43

4.2.2. Efficient Use 43

4.2.3. Roaming 43

4.2.4. Coordination 43

4.3. Base Station Siting 44

4.4. Backhaul 45

4.4.1. Technologies 46

4.4.2. Specific Examples: GigaBeam and Ceragon 47

4.4.3. Backhaul Bottom Line 49

5. EMERGING MOBILE INTERNET APPLICATIONS 50

5.1. Content Delivery 50

5.1.1. Example: Kindle 50

5.1.2. Benefits of Ubiquity 51

5.1.3. Innovation from Differentiation 52

5.2. Wireless VoIP 53

5.2.1. Example: Skype on the 3 Network 53

5.2.2. Example: Other Ways to Skype 55

5.2.3. Economic Implications of Skype 55

5.3. Wireless Video Streaming 55

5.3.1. Example Video Streaming System: MediaFLO 56

5.3.2. Example Video Streaming System: MobiTV 57

5.4. Mobile Augmented Reality 58

5.4.1. Example: Layar Platform 58

5.5. Internet of Things 60

6. MOBILE INTERNET POLICY ISSUES 60

6.1. Net Neutrality 60
6.1.1. Transparency-based Net Neutrality Lite 61
6.1.2. Transparency 62
6.1.3. Content Freedom 62
6.1.4. Application Freedom 62
6.1.5. Reasonable Limits on "Device Freedom" 64
6.1.6. Managed Services 65
6.1.7. Enabling Innovation 65
6.1.8. Specific Example: EU Telecoms Package 66
6.1.9. Services-Oriented Framework 67

6.2. Spectrum Policy 67
6.2.1. The Value of Open Spectrum 68
6.2.2. The Limits of Open Spectrum 68
6.2.3. What Licensed Spectrum Does Well 70
6.2.4. Spectrum Inventory and Review 70
6.2.5. Reassigning Spectrum 72
6.2.6. Future Proofing the Spectrum 73

7. POLICY RECOMMENDATIONS 75

7.1. Stick with Light-touch Regulation 76

7.2. Enact a Sensible Transparency Rule 76

7.3. Legitimize Enhanced Transport Services 76

7.4. Define Reasonable Network Management 77

7.5. Preserve Engineering and Operations Freedom 77

7.6. Review Existing Spectrum Licenses 78

7.7. Eliminate Redundant and Archaic Licenses 78

7.8. Protect Spectrum Subleasing 78

7.9. Cautiously Enable Secondary Uses 78

7.10. Allow the Experiment to Continue 79

List of Figures

Figure 1: Classical Internet Protocol Version 4 Header 17
Figure 2: Hot Potato Routing 22
Figure 3: Mobile Technology Summary 29
Figure 4: Carrier-sensing overhead in an IEEE 802.11g Contention System.................. 34
Figure 5: Carrier-sensing overhead in an IEEE 802.11n Contention System.................. 34
Figure 6: Deering's Hourglass.. 37
Figure 7: Internal Architecture of the RACS model 39
Figure 8: Cell tower disguised in cross by Larson Camouflage. 45
Figure 9: Internet is a Virtual Network composed of Many Physical Networks. 46
Figure 10: Cergagon FibeAir® IP-10 Switch 48
Figure 11: GigaBeam installation providing backhaul for Google's Mountain View Wi-Fi Network.. 49
Figure 12: Amazon Kindle... 51
Figure 13: 3 Skypephone 54
Figure 14: Example of FLO network 57
Figure 15: ATSC-M/H Simulcast 58
Figure 16: Beatles Discovery Tour starting at Abbey Road. 59
Figure 17: Licensed LTE is more efficient than license-exempt Wi-Fi. 71
Figure 18: Novatel Mi-Fi Device 74

1. Introduction

Ten years ago, the typical American used the Internet through a dial-up modem. Going on-line was a dramatic event accompanied by a full range of sound effects as the modem spat out a series of tones to make a connection and then exchanged screeches and whirs with the answering modem to assess the telephone network's signal quality. With luck, the network might support 48 Kbps downstream and 33 Kbps upstream. The Internet Service Provider (ISP) industry was still emerging, and more likely than not the early Internet consumer dialed-in to a walled garden system such as America On-Line or CompuServe. The primary application of the day was e-mail, but the adventurous explored Usenet discussion groups and tried this new thing called The Web. The Web was challenging because it didn't have a map, the pages were full of strange acronyms and opinions, and pictures dribbled onto the screen at a snail's pace. Surfing the web was like wandering the public library with your eyes closed and picking books off the shelf at random: always unexpected.

1.1. *Advent of Broadband*

The advent of broadband networking changed this system in many ways: it sped up the Web and brought indexers and mapmakers like Google and Yahoo! into the picture. It made e-mail a more useful, always-on system, and it changed the choice formula for ISPs. Instead of dozens of places to buy an equivalent low-speed service, we had a smaller number of broadband ISPs, but their service differences were real, and they actually competed on quality as well as price. Moreover, with the advent of broadband, the Internet began to create different kinds of applications, such as the Voice over IP (VoIP) systems from Vonage and Skype that lowered our phone bills and systems like Napster and KaZaA that magically provided us with free entertainment (we later found it was pirated, of course.)

Technically, it wasn't hard to bring VoIP to an Internet dominated by the Web. VoIP is a narrowband application that scarcely consumes more bandwidth than a dialup modem. The technical demands of web surfing are greater – all those pictures to download – but web surfing is on-again, off-again from the network's point of view. Web pages require human time to read, and while that's going on, the network has capacity to spare. Adding VoIP to this system was just like pouring sand into a bucket of rocks. There was plenty of room to spare as long as we weren't too carried away with the free entertainment on The Pirate Bay.

From the consumer's point of view, the transition from dial-up to broadband was perfectly seamless. With broadband the web got faster, applications became more enjoyable, and eventually the Internet became more or less indispensible, despite the nuisance of spam and the occasional virus. We were no longer locked-in to the small community on AOL; we could be inspired as well as irritated by people all over the world and we had much of the world's vast stores of information, commerce, learning, and cultural heritage at our fingertips.

1.2. *Rise of the Cell Phone*

While the Internet was permeating modern life, a parallel development was taking place that would have perhaps even greater significance for billions of people all over the world. On April 3, 1973, Martin Cooper, the general manager of Motorola's Communications Systems Division, had placed the first telephone call ever from a portable cell phone. By the turn of the century, cell phones were common business tools, and they eventually became the preeminent global means of personal communication at a distance. For billions of people in the undeveloped world, the cell phone was the first telephone they ever had, and it quickly became the indispensible means of communicating.

Inevitably, these two transformative technologies began to merge, giving rise to an ocean of social and economic benefits and to a host of policy challenges. The Internet had a legacy of distributed computers, open systems designed around *end-to-end arguments*, a reflection of its heritage as a tool originally built to stimulate academic research in the new communications technology known as packet switching.[2] Cellular telephony had a more humble legacy, as it simply aimed to extend the reach of the existing telephone network, not replace it wholesale with a bombproof alternative.

1.3. *Convergence*

From the engineering viewpoint, the cell phone network and the broadband Internet could hardly be more different: one is mobile, the other locked-down in space; one is high capacity, the other narrowband; one is personal and intimate, involved in relationships where "content" doesn't exist until the moment it's communicated, the other is part of a wide-open, always-on system that pulls information from multiple sources at once, and one is built around portable battery-powered devices, while the other draws power from a plug.

People now want both the Internet and mobility, so it became necessary to bring the Internet to the cell phone, just as it had once been brought to the home phone, and vice versa. The mobile Internet borrowed some of Steve Jobs' pixie dust and transformed the cell phone into a smartphone, and then expanded the mobile network from narrowband to broadband. It's now beginning to rebuild the Web into a personalized, real-time system that responds to the locations, tastes, and whims of billions of people as they live their lives, at their desks or moving through the world with the flow of experience.

Even with all of network magic we enjoy today, we're still at a very early stage in the development of the Mobile Internet; with any luck, we'll one day look back to where we are today the same way we remember the dial-up era, wondering how we could ever have been so naïve as to tolerate the limitations of the bygone era. The flowering of the Mobile Internet will only come to pass, however, when engineering and policy collaborate to successfully overcome the major challenges standing in the way of the development of a Mobile Internet that lives up to its full potential. For this to happen, policymaker must refrain from strangling the Mobile Internet with excessive regulation.

1.4. *Reader's Guide to the Report*

This report consists of two major parts, the first on technology and the second on policy. The technology section is a deep dive into the history and development of both the wired Internet and the mobile network, ending in an explanation of the way the mobile network connects to the Internet. The information in these sections informs the policy discussion that follows by showing implications that policy choices have on technical evolution. At the conclusion of the technology section, the patient reader is rewarded with a palate-cleansing glimpse at new and emerging applications before the report turns fully to policy matters.

One of the challenges the Mobile Internet faces is the reconciliation of the norms of two technical cultures that have always seen the world in different ways. C. P. Snow was much too optimistic when declared there was a single technical culture; in fact, with respect to the mobile Internet, there are two. The other policy challenge relates to radio spectrum, which has variously been called the lifeblood, the oxygen, and the energy of the Mobile Internet. The report concludes with a series of recommendations for policymakers on the key issues before them in the United States.

2. Issues in Internet Design and Operation

There's a tendency to view the Internet as a force of nature, something that sprang up spontaneously. In fact, it's a man-made system was designed by engineers in a cultural context that could easily have been designed differently. The different orientations of the Internet's "Netheads" and the telephone network's "Bellheads" are the source of much unnecessary conflict. Oftentimes it seems that members of these tribes disagree with each other for the sake of it, but most of their engineering differences are related to the relative importance of different kinds of applications within their networks.

2.1. *Brief History*

The Internet was conceived in the early- to mid-1970s to interconnect three research networks: ARPANET; the San Francisco Bay Area Packet Radio Network (PRNET); and the Atlantic Packet Satellite Net (SATNET).[3] By 2010 standards, these constituent networks were very primitive; each was the first of its type and computer technology was much less advanced than it is today. Each was built on a different technology, and each was separately administered.

As the operational parameters of these networks varied radically, designers of the Internet protocols, led by ARPA's Bob Kahn and Vint Cerf, couldn't rely on network-specific features to pass information between networks; instead, they adopted a lowest common denominator approach for the Internet Protocol (IP), the "datagram" network abstraction borrowed from France's CYCLADES network.[4] In order to match the speeds of sending and receiving stations (called "hosts," following timesharing terminology), the designers developed a sliding window overlay above IP called Transmission Control Protocol (TCP) that conformed to the model established by the CYCLADES Transfer Station protocol.[5]

IP is very simple, and is in fact is more a format than a protocol since it doesn't describe any specific sequences of behavior; it's easily implemented over any packet-switched network. TCP, on the other hand, is a complex, high performance system that can keep multiple packets in flight between source and destination, a crucial requirement of high delay satellite networks. TCP is easily an order of magnitude more complex than the rudimentary end-to-end protocols of the day such as IBM's Binary Synchronous Communications and ARPANET's Network Control Program.

The Internet design team allocated functions as they did in order to provide the greatest opportunities for experimentation with network protocols. Subsequently, researchers developed *end-to-end arguments* aligning such function placement with a general theory of distributed system design, and in so doing inadvertently generated elements of the policy argument that has come to be known as network neutrality. End-to-end in the hands of the policy community has very different implications than it has in the engineering world.[6]

2.2. *Modularity*

The Internet is a collection of separate, replaceable elements called "modules" by engineers; its overall structure has been described by philosopher David Weinberger as "small pieces loosely joined".[7]

> Internet modularity preserves the opportunity for experimentation on applications that take the Internet as a platform, and on the elements of the Internet itself, such as TCP, IP, and the systems of naming, routing, and security.

Modular design decomposes a technical system into functions that can be implemented in separate components called "modules". "Platforms" such as the Web are collections of modules. This strategy, sometimes called "divide and conquer," has a number of benefits. It enables functions to be specified, developed, and tested in isolation from the rest of the system, facilitates reuse of modules in different systems, and makes it easy to improve the implementation of a function without destabilizing the entire system.

Modular design is practiced in a number of technical fields, including computer science and network engineering. As the seminal work on modular computer systems design was presented at the same forum as the seminal work on the design of ARPANET, it's fair to say that modular computer systems and the Internet developed hand-in-hand.[8]

Internet design begins by distinguishing internetworking from networking, thus confining the Internet's role to interconnecting networks rather than providing basic network services. This important element of Internet design often escapes policy advocates, who mistakenly believe that the Internet specifies a particular method of network operation. An Internet is a virtual network (or a "meta-network") that works with networks as they are, imposing a minimal set of requirements.

The Internet doesn't care whether a member network is public or private, fair or unfair, fast or slow, highly reliable or frequently broken. The operator of an Internet member network may route all packets on equal terms, and he or she may differentiate.

IP was initially designed to preserve service information that may have had no meaning outside a particular member network, such as the Type of Service specified in bits 8 – 15 of the IP header and subsequently redefined by the Differentiated Services protocol.

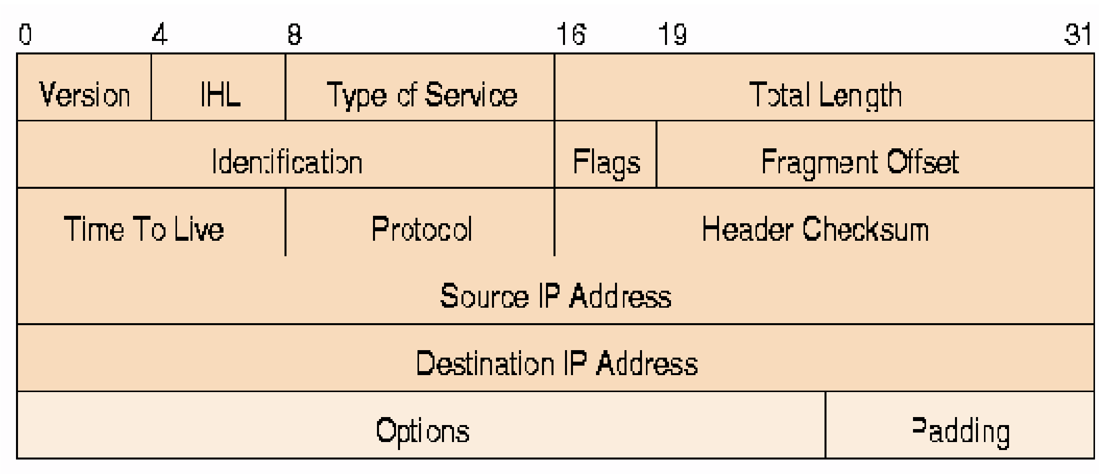

Figure 1: Classical Internet Protocol Version 4 Header

RFC 795 specified the interpretation of the Type of Service field by ARPANET, PRNET, SATNET, and AUTODIN II.[9] Because the Internet design delegates such matters as service differentiation to physical networks, a myth has developed to the effect that the Internet is a "stupid network" that can't differentiate. In fact the Internet leaves all functions not directly pertinent to cross-network packet formatting and payload processing to individual networks; the Internet is not hostile to service differentiation, but differentiation is outside the scope of internetworking.

Modular design separates functions and creates design hierarchies. Modular systems such as the THE multiprogramming system and the Internet protocols organize vertically, into higher-level and lower-level functions, where dependency increases with altitude.

The Internet's modular design produces benefits for innovation by creating platforms that simplify application development. Just as physical networks are platforms for IP datagrams, IP datagrams are a platform for TCP, which in turn is a platform for the web, which is a platform for Facebook, which serves as a platform for Facebook applications. Each platform simplifies the creation of new applications by managing aspects of the application, but this simplification comes at a cost in terms of efficiency.

Internet modularity preserves the opportunity for experimentation on applications that take the Internet as a platform, and on the elements of the Internet itself, such as TCP, IP, and the systems of naming, routing, and security.

2.3. *Efficiency*

The design of the Internet generally sacrifices efficiency to flexibility, as one would expect in a research network.

The separation of functions required by modular design tends to reduce system efficiency by partitioning information, increasing generalization, and imposing interface costs. An application that relies on a platform function for authentication, for example, has to request and wait for authentication services from the platform. As the platform is more general than the application, its way of performing authentication may require more processing time than the application would require if it performed this task itself; the benefit is that the application programmer doesn't need to worry about authentication and can focus on the more unique aspects of application logic.

Modular organization lends generality to systems, simplifying higher-level components, but in so doing increases the information-processing burden. In most cases, this is a reasonable tradeoff: skilled system programmers are highly paid, and hardware components are cheap; modularity reduces the number of software bugs; and modular systems can re-use components developed and verified in other systems. In cases where efficiency is a paramount goal, modularity can be a burden if the system is not so well designed that modules are partitioned exactly according to natural boundaries.

Mini-Tutorial: Why is Internet downloading faster than uploading?

Each of the 1.6 billion Internet users in the world relies on fewer than 1,000 Internet Exchange Points or IXPs to get from one network to another. Between the consumer and the nearest IXP are a number of switches that "aggregate" or combine packets sent on lower speed data links onto higher speed data links. In the opposite direction, each of these switches "disaggregates" or un-combines. The number of times a wire can be aggregated is limited by the speed of the fastest technology the IXP can buy, and by the number of networks the IXP can interconnect. Currently, most IXPs interconnect ISP networks at 10 Gbps. Upload speed is therefore limited by the Internet's traditional methods of operation. High-speed Content Delivery Networks don't aggregate as much as large ISP networks, so their upload speeds are faster than those of ordinary consumers.

There are many examples of the Internet's inefficiency in action. The best well-known concerns the Internet's congestion control system, implemented between interior IP routers and end-user systems.[10] This system requires new connections to begin in a low throughput condition called "slow start," defeating the desire of applications to transfer information quickly. It also contains a feature called "multiplicative backoff" that divides the application's self-imposed bandwidth quota in half in response to each indication of congestion. The net result of these features is to prevent Internet core data links (communication circuits) from utilizing more than 30 percent of designed capacity.[11] Given that Moore's Law has caused data links to become cheaper and faster since this system was deployed, its benefits in terms of stability outweigh its inefficiency. The same calculus does not apply to wireless data links, however.

2.4. *Manageability*

The Internet is relatively weak in terms of manageability, in contrast to its constituent physical networks. ARAPNET was managed from a central control point in Cambridge, Massachusetts, where operators were able to view the status of each separate data link from a single console, as PSTN operators can do. The Internet standard for network management, the Simple Network Management Protocol (SNMP) relies on a system for viewing and modifying the state of physical network components that hasn't responded well to security challenges or the sheer growth in the Internet's size, although it has proved helpful in tracking down cybercriminals in some cases.[12] SNMP is dependent in any case on physical network facilities. In response to SNMP's weaknesses, the members of the Broadband Forum (mainly network operators) have devised an alternative system more aligned with the security norms of telephone network management, TR-069.[13]

The elements of the Internet unique to internetworking, principally routing and the information exchanges that make routing possible, are handled by IETF standards such as Border Gateway Protocol (BGP). BGP was thrown together hastily in order to allow the transition of the Internet from the subsidized NSFNET backbone to the current system where for-profit entities provide network interconnection and packet transit. While BGP had the ability to assign and exchange QoS information in routes, it was very rudimentary due to unresolved research questions and "political" issues.[14] Fundamentally, the problem of network-wide QoS was complex and network operators were unmotivated to solve it in 1989 absent a compelling application need. Lackluster support for QoS routing and the predominance of a single application slowed the development of large-scale QoS across the public Internet, which is one reason that net neutrality advocates insist that the Internet is a "best-efforts network." BGP doesn't alter the nature of IP, however, and there are ways around the initial limitations in BGP regarding QoS.

Mini-Tutorial: Is the Internet a first-come, first-served network?

Many advocates of anti-discrimination regulations insist that the Internet has always handled all packets on a first-in, first-out basis. This common simplification has never been true. Internet edge routers, the devices that connect ISP networks to the Internet core, are typically configured to apply a "weighted fair queuing" algorithm across either packet streams or users to ensure fair and equal access to common resources. Simply put, fair queuing systems select packets from each user in round-robin fashion. Advocates of the first-in, first out rule confuse the reality of network management with the simplified public story.

BGP is in crisis, according to a report issued by the IETF's Internet Architecture Board Workshop on Routing and Addressing in 2007.[15] The transition to IPv6 and the onrush of new mobile users place requirements on Internet edge routers that exceed the pace of progress in fundamental computer hardware technology.

The Internet architecture and protocols also lack direct facilities for dealing with malicious behaviors such as bandwidth hogging and Denial of Service attacks, which it relegates to network operators, IP routers, firewalls, and Network Address Translators. The architecture needn't address these issues since the constituent networks are perfectly capable of handling them on their own. Some advocates insist that the Internet's architecture makes "discrimination" impossible.[16] It's difficult to see where this naïve idea comes from, since every IP datagram displays source and destination IP addresses, and IP's most common payload, TCP, prominently displays a destination port number clearly identifying the application protocol.[17] Most IP payload is carried in clear text, so this information is discernable to anyone with access to a shared network link and an off-the-shelf protocol analyzer such as Wireshark. The clear-text IP format is practically an invitation to discriminate.

It's likely that significant changes are coming to the Internet in order to improve basic manageability, especially in the routing function, so historical limitations in BGP shouldn't drive the network regulation debate one way or another. Moreover, advances in congestion management are likely to connect economics with the resolution of micro-congestion events. This is a reasonable approach for a user-financed Internet.

2.5. *Innovation*

The Internet has unquestionably served as a tremendously successful platform for innovation. Economic powerhouses such as Google, Amazon, Yahoo!, eBay, Facebook, Twitter, Skype, and YouTube are household names thanks to the Internet, and some have become verbs in official dictionaries.[18] Successful Internet innovations are typically file transfer-oriented web applications.[19] Even video streaming, the Internet's television analogy, is implemented on YouTube and Netflix "Watch Instantly" as a file transfer, which is why traditional "trick play" VCR features such as fast forward and rewind function so poorly on these systems. Communication-oriented innovations such as Skype and other VoIP services don't follow the file transfer paradigm, but their share of the innovation space (as well as their contribution to total Internet traffic) is relatively small.

Web applications have predominated as long as the Internet has been a mass phenomenon, so the innovation barrier that new systems have had to overcome is co-existence with the web. This hasn't been difficult. Like sand poured into a bucket of rocks, VoIP uses the capacity that the web can't consume because of the on-again, off-again "episodic" nature of web access.

Because it takes human time to read web pages, the web generates short periods of network activity followed by long periods (in computer terms anything over a millisecond can be considered "a long time") of inactivity. VoIP is an adaptable, persistent, narrowband application that can run with an average allocation of 4 to 100 kilobits per second; it easily finds transmission opportunities on all but the most overloaded broadband facilities.

VoIP didn't have any major problems on the Internet until a second non-web category of innovation emerged, peer-to-peer file transfer. Peer-to-peer applications such KaZaA and BitTorrent, used mainly for piracy,[20] have an appetite for bandwidth that exceeds that of VoIP and the web by several orders of magnitude: The typical web page is 130 kilobytes, while the typical pirated video ranges from 350 to 1,400 megabytes per hour, depending on resolution.[21] The typical peer-to-peer transaction is equivalent to loading two web pages per second for an hour or two.

Peer-to-peer also fills the spaces between the Internet's rocks – the idle periods between web accesses – and makes VoIP a challenging application, and it continues to do so after the user has downloaded a particular file since it is both a file server and a download client. Historically, the Internet has relied on the good will and good behavior of end users to prevent the instances of congestion that can cause applications to fail, but peer-to-peer design chooses not to comply with these voluntary norms of conduct.[22]

As innovation seeks opportunities beyond the web paradigm and applications diversify, we should expect to see more instances of friction between applications such as Skype and BitTorrent. Some advocates who looked backwards can try to pretend that these conflicts are not real, but they are and they will test the ability of the regulatory system to respond effectively, and its record thus far does not inspire confidence. For example, the FCC's ruling on petitions filed against Comcast by peer-to-peer indexer Vuze, Inc. and a host of law professors and public interest groups in 2007 was troubling on both factual and legal grounds, and is likely to be overturned by the courts.[23]

Moreover, on today's Internet, most innovation friction takes place between some applications and other applications, not just between applications and network operators. The net neutrality movement's exclusive focus on operator behavior obscures this very important fact

2.6. *Subsidies*

The Internet's current business models effectively subsidize content-oriented innovation. Packet routing follows a model that assigns the lion's share of information transfer costs to the ISPs that are content receivers, rather than content creators and middlemen. As IETF member Iljitsch van Beijnum explains:[24]

> Unless you go out of your way to make things happen differently, Internet routing follows the early exit or "hot potato" model: when traffic is destined for another network, it gets handed over to that network as quickly as possible by sending it to the closest interconnect location.

When ISP executives such as Ed Whiteacre, the former CEO of AT&T, complain about innovators "using [ISP] pipes for free," hot potato routing is part of the context, because handing a packet from one network to another also hands over the costs of transporting it the rest of the way.[25] When packets are handed over as early as possible, hot potato style, the receiving network ends up paying the bulk of the costs for end-to-end packet transport.

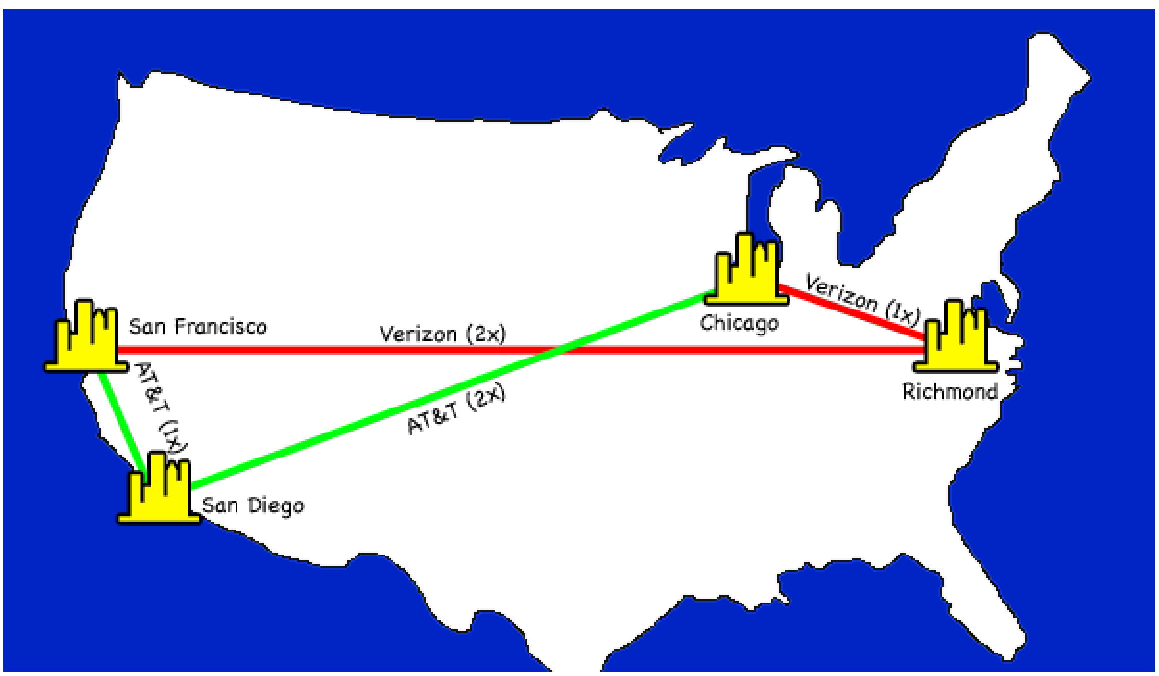

Figure 2: Hot Potato Routing

The network diagram in Figure 2 helps us understand hot potato routing between the Verizon customer in Richmond and the AT&T customer in San Diego. Packets sent by the Richmond customer leave the Verizon network in Chicago, and are transported most of the way by AT&T. Packets sent by the AT&T customer leave the AT&T network in San Francisco, and are transported most the way by Verizon. For streaming applications such as YouTube and Netflix, 99% of the traffic goes from server to user, so the receiving user's network pays most transit costs. The question of surplus doesn't arise until transit costs are covered.

When we combine the practice of hot potato routing with the fact that web users receive much more data than they transmit, by a factor two or more orders of magnitude, it becomes clear that ISPs and their immediate customers do in fact pay most of the costs of transporting packets across the Internet. In this sense, it's somewhat analogous to how consumers pay the Post Office to have a package shipped to their house that they bought online or in a catalog: on the Internet, users "order packets" from other places and pay for most of their delivery. Whiteacre's complaining notwithstanding, this is not a bad system.

This economic system has unquestionably helped enable the Internet innovations with which we're all familiar by lowering entry barriers to small, new content creators and aggregators. Preserving this financial system is perhaps the central issue for network neutrality advocates who characterize deviations from it as a counter-productive wealth transfers from innovators to incumbent operators. As Institute for Policy Integrity economists J. Scott Holladay and Inimai M. Chettiar say:[26]

> At its heart, net neutrality regulation is about who will get more surplus from the Internet market. Retaining net neutrality would keep more surplus in the hands of the content providers, and eliminating it would transfer some surplus into the hands of the ISPs. Changing wealth distribution would affect the ability and incentive of the respective market players to invest in the portions of the Internet they own.

Making retail Internet customers cover most of the costs of generic IP transit isn't the end of the story, however. When IP transit services are completely decoupled from content, applications, and services, the economic incentives of network operators become especially misaligned with those of the innovators whose systems depart from the traditional web model. If we want next-generation networks (NGN) to enable next-generation applications, we probably need next-generation economics to align incentives, by providing senders of packets (e.g., application and content providers) with the ability to pay a bit more to send packets by "Express Mail" rather the traditional best-effort "First Class mail." The complete decoupling of applications from networks creates regulatory friction and the need for enhanced packet delivery systems.

Mini-Tutorial: How does the Internet adapt to all these different networks?

The Internet's secret is that its smallest and most fundamental element is nothing more complicated than a message format, which we call "Internet Protocol." Protocols are usually complicated procedures, but IP is simply a way of organizing messages so that they can pass between networks. The U. S. Postal Service can pass postcards to Canada Post because both postal services share a common understanding of addressing (and revenue sharing,) not because they operate the same trucks, trains, and airplanes.

However, the subsidy for basic packet delivery that lowers entry barriers for small businesses evaporates for many larger ones. Most large content-oriented firms use for-fee Content Delivery Networks (CDNs) to deliver packets to ISPs from nearby locations with minimal transit expenses and faster and more reliable delivery. Even though the large firm's path to the ISP is short and relatively inexpensive, the cost of transporting a single copy of a movie to a CDN and then paying a higher fee for each copy the CDN delivers to the ISP is worthwhile from the Quality of Service (QoS) point of view. CDNs decrease ISP transit costs and reduce the traffic load on the Internet core, so they represent the happy alignment of interests that can come about when parties voluntarily choose to abandon subsidized transport for enhanced service: quality improves, costs are realigned, and the core becomes more robust. These things are true when the ISP provides the CDN service as well.

An additional complication in the existing system of allocating payments on the Internet arises in the so-called "middle mile" networks between ISPs and Internet Exchange Providers (IXPs), especially in rural settings. As Christopher Marsden explains, Internet regulators have long overlooked the economic role of the middle mile: "until recently, many analysts did not fully appreciate that so much traffic on the Internet was monetized and had to pay its way."[27] High transit fees are the reason some small ISPs (e.g. Wireless ISP LARIAT in Laramie, Wyoming) ban the use of P2P outright on their lowest-tier pricing plans.

2.7. *Specific Example: Inter-Domain Quality of Service*

Quality of Service (QoS) is the technical term for systems that match application requirements to available network capabilities. Some applications, most notable telephone service, require low delay or latency between on caller and another, and other applications simply require low cost. QoS differentiation was part of the original design the Internet Protocol and of subsequent work by the Internet Engineering Task Force on the following standards:

- Multiprotocol Label Switching (MPLS)[28]
- Differentiated Services (DiffServ)[29]
- Integrated Services (IntServ)[30]
- Real Time Protocol (RTP)[31]

- Real-Time Streaming Protocol (RTSP)[32]
- Resource Reservation Protocol (RSVP)[33]
- Work in progress on Congestion Exposure
- Work in progress on Inter-domain routing

In addition, ICANN has assigned numbers to populate portions of BGP Community Attributes with QoS level identifiers. The idea that the Internet requires a single service level persists because of the confusion between theory and practice, architecture and implementation.

Recall that BGP was created in a few short months out of the necessity of replacing the NSF backbone, at a time when nearly all Internet use was file transfer oriented. BGP is a mechanism that allows network operators to exchange routes based on policy, such that each partner network of a given network operator can be shown a different set of routes tailored to the business relationship between the networks. BGP is capable of advertising routes based on a variety of internal policies, which may reflect QoS attributes. In practice, BGP has not been used for this purpose simply because networks have not chosen to include QoS routing in their terms of interconnection. The ability to do this has always been present, but the will has been lacking for a number of reasons.

For one thing, there's little to be gained in applying QoS on the high-speed links (40 and 100 Gbps) that form the Internet's optical core. Active QoS measures are most valuable on data links where load is highly variable, but the core is so far removed from statistical variations and so well provisioned for peak load that QoS measures would rarely be invoked. Core traffic is aggregated from so many different sources that its contours only show slow diurnal variation. The value of QoS increases, however, as the packet traffic on the highest speed links is disaggregated onto slower links with more variable load. The problem with implementing QoS at this level, between origin network and destination network, is complicated by the fact that the operators of the relevant networks may not have a business relationship with each other. Consider two networks, A and B, who exchange packets through a common transit network, C. A and B both have business relationships with C, but not with each other. A and B have no reason to specify QoS with core network C, for the reasons given. Hence, they have no agreement with each other for QoS.

Developing QoS relationships between non-core networks would require a number of negotiations to take place that haven't been necessary in the past. Core networks – both Tier 1s and Tier 2s – can facilitate this process by adding QoS to the agreements they have with other, as some have done according to the cryptic information that's publicly available about peering agreements.[34] What's needed is a set of transitive[35] agreements on core networks that transfer to edge networks automatically, essentially a marketplace like the stock exchange or eBay that allows the owners of the 30,000 networks within the Internet to make such deals.

Another reason that end-to-end QoS isn't widely implemented is that BGP is a fundamentally insecure protocol. Network operators misconfigure BGP route advertisements on a regular basis, and a number of attacks are trouble for BGP.

> The hijacking of YouTube's routes by the government of Pakistan was only possible because the Internet is fundamentally insecure.

The best-known example of this was the misconfiguration of routes to YouTube by a network technician in Pakistan intended to block access from within that country. YouTube had a block of IP addresses consisting of a 23-bit network number and 9-bit host numbers. The Pakistani technician set his router to advertise two 24-bit network numbers for YouTube, which allowed his equipment to block requests to YouTube (network administrators call this practice "blackholing.")

BGP prefers more specific routes over general ones, so as soon as the bogus Pakistani routes were advertised across the Internet they became the preferred routes for everyone; a 24-bit network address, even if it's fake, is more specific than the 23-bit address advertised by YouTube, even though it's real. Making QoS a feature of routes would make it that much easier for malicious users or careless administrators to exhaust network resources by elevating P2P to highest priority, for example. This would be disastrous for the victim network. BGP insecurity is a second barrier to wide-scale QoS deployment through the Internet core.

Nevertheless, it is a fact that most home gateways support user-controlled QoS, as do all major operating systems, as well as the Internet routers sold by major firms such as Cisco, Juniper, Huawei, and Alcatel. Hence, the barriers to the implementation of end-to-end QoS across the Internet can be overcome. Ultimately, it's an operational and business case problem that can be addressed by ISPs when they see the need, so long as regulators haven't foreclosed the option and an efficient marketplace exists in which these transactions can be made.

2.8. *Why the Internet Works*

According to Professor Mark Handley of University College, London, the Internet "*only just works*."[36] It has been architecturally stagnant since 1993 because it's difficult to make changes in such a large system except as they're motivated by fear:

> …technologies get deployed in the core of the Internet when they solve an immediate problem or when money can be made. Money-making changes to the core of the network are rare indeed — in part this is because changes to the core need to be interoperable with other providers to make money, and changes that are interoperable will not differentiate an ISP from its competitors. Thus fear seems to dominate, and changes have historically been driven by the need to fix an immediate issue. Solutions that have actually been deployed in the Internet core seem to have been developed just in time, perhaps

> because only then is the incentive strong enough. In short, the Internet has at many stages in its evolution only just worked.

IP Multicast, Mobile IP, Quality of Service, Explicit Congestion Notification, secure BGP, and secure DNS are all significant enhancements to the Internet architecture that solve real problems and have not been widely adopted (although we are finally making progress with DNS.) One of the implications of the Internet's end-to-end architecture is that major changes need to be implemented among all, or at least most of the hundreds of millions of end-user computers and network routers that comprise the Internet. Greed alone has never been sufficient to motivate such change, only collapse or near collapse has been a sufficient. The Internet itself only replaced ARPANET's NCP protocol because ARPA issued an edict that users had to upgrade to TCP by January 1, 1983 or lose access to the network.

While network neutrality advocates fear change in the Internet, engineers fear the lack of change:[37]

> …there has been no substantial change at layer 3 [IP] for a decade and no substantial change at layer 4 [TCP] for nearly two decades. Clearly then the Internet is suffering from ossification. The original general-purpose Internet which could evolve easily to face new challenges has been lost, and replaced with one that can only satisfy applications that resemble those that are already successful…

> … The number of ways in which [the Internet] only just works seems to be increasing with time, as non-critical problems build. The main question is whether it will take failures to cause these problems to be addressed, or whether they can start to be addressed before they need to be fixed in an ill co-ordinated last-minute rush.

Primarily, the Internet works because network operators spend enormous amounts of money to ensure it works as well tomorrow as it did yesterday when the workload was less. It doesn't solve any problem – other than file transfer – especially well, but it solves many problems just well enough for general utility. The backlog of non-deployed enhancements and upgrades is growing, as is the friction between new devices, new applications, and new users and the installed base.

The most immediate stimulus for a wholesale upgrade of the Internet architecture – the only thing that might motivate a major improvement in the design of TCP and IP – is address depletion, using up all possible IPv4 addresses. Unfortunately, the large-address replacement for IPv4, IPv6, does not sufficiently address the myriad of small but growing problems that hold convergence back. IPv6 deployment will cause a crisis in edge router design because it will cause a massive increase in the number of routes each edge router must hold in its specialized memory, however, and the resolution to that problem may enable necessary changes to be made.[38] Opportunities to upgrade the Internet's major protocols – BGP, IP, and DNS – are few and far between, so they can't be squandered.

It would be unfortunate if well-meaning regulators added to the Internet's general ossification by forbidding necessary and constructive changes while the question of the Internet's future architecture remains unresolved.

3. Issues in Mobile Network Design and Operation

While the Internet was designed to support a simple group of low duty cycle applications (remote login, e-mail and occasional file transfer) for locked-down computers shared by a modest number of highly-skilled, trustworthy users in a non-commercial setting, mobile telephony's heritage is one of real-time communication-oriented applications, large numbers of mobile users, and personal devices competing for market share. It's not surprising that there's friction between Internet culture and mobile culture.

3.1. *Brief History*

The mobile network was originally designed to serve as an extension of the telephone network that added mobility at the network edge without altering the telephone network's fundamentals. Initially, it used analog technology, and converted to digital in the 1990s. Data services were a special feature added on to the mobile network roughly during the period of its transition from analog to digital. As presently operated, the mobile network is still more efficient at providing telephone service than data service.

Data services on mobile networks were initially enabled by the Cellular Digital Packet Data (CDPD) standard developed in the early 1990s. CDPD provided data transfer rates of 9.6 to 19.2 Kbps, roughly comparable to PSTN modems; it was adopted by AT&T, Cingular and Palm. Primarily noteworthy for providing digital service over the analog cell phone networks of the day, CDPD demonstrated the utility of data over mobile networks. It was subsequently replaced by faster, all digital systems such as EDGE, WCDMA, HSPA and HSPA+; the next big phases of mobile networking are LTE and LTE Advanced, data-oriented architectures. When mobile networks convert to LTE, the transition from a voice-centric to a data-centric network will be complete.

	3GPP version	Year	Coding	Mod-ulation	MIMO	Channel Width in MHz	Peak D/L Mbps	Typ-ical D/L Mbps	Efficiency bits/Hz	Delay ms.
EDGE		1997	TDMA	QPSK		5	0.5			500
WCDMA	Rel 99	1999	WCDMA				1.8			150
HSDPA	Rel 5	2002					3.6			100
HSUPA 7.2	Rel 6	2004					7.2	0.7-1.7		75
HSPA 14.4				16 QAM			14.4			
HSPA+ 21	Rel 7	2007		64 QAM			21.6	1.5-7		
HSPA+ 28	Rel 7				2x2		28			
HSPA+ 42		2009				10	42.2			
HSPA+ 84		2010			2x2	10	84			
LTE	Rel 8	2009	DFTS/ OFDM		2x2	1.4 - 20	150	4-24	1.69	25
					4x4		300		2.67	
LTE Advan-ced	Rel 10+	2011	OFDMA/ SCDMA		8x4	10 - 100	1000	100	3.7	< 25

Figure 3: Mobile Technology Summary[39]

The data rates of these systems have doubled roughly every 30 months, as predicted by Cooper's Law.[40] By way of contrast, Butter's Law predicts that the data rate of optical fiber doubles every nine months; these rates of improvement are different because they reflect the differential effects that Moore's Law has on different semiconductor processes. Moore's Law affects mass-market chips built in the newest fabs, but many radio parts are analog components whose manufacturing process progresses more slowly. It's also the case that air is a more challenging medium for data communication than copper or optical fiber; some have said that wireless is a generation behind wired systems and always will be.

This is important because the rate of progress for Internet applications is largely driven by price/capacity improvements in physical networking since the virtual network – the Internet protocols – has been stagnant since 1993. As Internet use shifts to wireless networks with slower intrinsic rates of advance, we might expect a slower overall rate of innovation, assuming all other factors remain constant. We should also expect increased fracture between mobile applications and stationary ones, in part because the bandwidth gap between the two can only grow larger but also because of intrinsic differences between fixed and mobile applications. Such factors are among the causes of application diversity. These factors combine in such a way as to suggest that innovation will actually accelerate.

3.2. *Modularity*

The system architecture of wireless networks differs substantially of from that of wireline, Internet-oriented networks in two major ways. Obviously, mobile networks support mobility, so users don't need to be attached to a wire to be connected. Second, mobile networks require subscriber identity; hence they distinguish users and devices from their points of attachment to the network. The Internet Protocol, on the other hand, only identifies devices by their points of attachment, and fails to identify users at all. The IETF has devised two protocols meant to add mobility to IP, Mobile IPv4 and Mobile IPv6, but neither has been widely adopted because of overhead and functional shortcomings, most importantly regarding security and packet routing overhead.

Mobile IPv4 routes first to a constant home location, and from there takes a second route to the current location, all at the expense of latency, which is unfortunate for voice communication (voice packets must arrive within 150-200 ms in order to please the ear and the brain, and doubling routing overhead doesn't do them any good.) Mobile IPv6 resolves this problem, but any form of route optimization on the Internet raises security concerns, as the Internet lacks a comprehensive security architecture.

Mobile networks contain one set of modules that roughly correspond to the elements of networking defined by the Open Systems Interconnect Reference Model (OSI) reference model at the Data Link Layer and at the Physical layer, such as the GPRS Core Network, the UMTS Terrestrial Radio Access Network (UTRAN) and the UMTS Terrestrial Radio Access (UTRA) air interface. Unlike wireline networks, mobile networks include additional functions such as the Mobile Application Part (interface for mobile-to-mobile calling and SMS), mobility routing, a PSTN interface, subscriber identification, and diverse application support within the network itself.

The newest wireless technology, called Long Term Evolution (LTE) replaces the current GPRS Core Network with an IP-based system called System Architecture Evolution (SAE), which will enable most of the traditional wireless device support functions to be provided by IP-based applications inside or outside of the carrier network. LTE is more than a fast radio; it's a wholesale redesign of mobile network architecture. The GPRS Core Network is a system designed for telephony that can do a bit of packet data, but SAE is a pure packet system that can do a bit of telephony.[41] In other words, the demands of the Mobile Internet are forcing the internals of the wireless network to become more like those of the wired Internet.

The rate of change in system-wide architecture is much faster in mobile communication networks than in the Internet, despite higher system complexity, greater hardware constraints, and a much larger user base. The transition to LTE is comparable in scope to the transition from ARPANET to the Internet in the early 1980s: a major reorganization that makes the system less expensive and more flexible in the long run. However, the LTE transition will be accomplished on a network of five billion users, while the Internet transition affected no more than a few thousand.

3.3. *Efficiency*

Cooper's Law increases in the rate of radio-borne data streams come about from increases in efficiency in all parts of the wireless network: better antennas, more effective coding of digital information onto analog radio waves, and better Medium Access Control (MAC) protocols. In the OSI Model, these functions all take place "underneath" IP, but many functions of radio networks take place in the service space "above" IP, particularly those that relate to the origination and termination of phone calls and to the interface between the mobile network and the fixed-line Internet. LTE features innovation in both the radio network and the service space (the part of the system that provides telephone, data, and video services.)

Mini-Tutorial: Why is Contention less Efficient Than Scheduling?

License-exempt "Open Spectrum" wireless systems such as Wi-Fi typically employ a contention protocol to determine which station gets access to the shared radio channel at any given time. Contention protocols require stations to sense "dead air" for an interval of time before beginning a transmission. The interval must be long enough to allow for propagation across the entire network. Propagation time is independent of data rate. As data rate increases, the length of time required to transmit the typical packet decreases. Hence, the higher the data rate, the greater the overhead of the dead air interval. In 802.11g Wi-Fi systems, contention overhead is close to 50% on the best case, and for 802.11n Wi-Fi, it can exceed 80%. For this reason, high-speed wireless systems cannot use contention protocols.

Radio network improvements begin with the use of Multiple-Input, Multiple-Output (MIMO) antenna technology. This innovation, which is also used in the IEEE 802.11n Wi-Fi standard, allows radio receivers to capture and process information generated by radio noise such as the reflected multipath radio signals that non-MIMO systems filter out as noise and reject. It does this by attaching multiple radios, each with its own antenna, to a single handset or base station, and separately processing and then combining the data streams received by each radio. In theory, MIMO multiplies wireless system capacity by the number of radios deployed. MIMO systems are identified by the number of transmit and receive radios used; a 3 x 2 MIMO system would have three downstream radios and two upstream radios, for example. In principle, MIMO systems may benefit from the use of as many as 64 radios, but due to power limits and diminishing returns, LTE assumes a baseline 2 x 2 configuration.

At the level of coding, LTE uses a scheme known as Orthogonal Frequency Division Multiplexing (OFDM). OFDM works very well on MIMO systems and is particularly resistant to common types of noise found in mobile settings. Most previous advances in radio data rates have come about due to advances in coding that increase the number of bits of information conveyed by a single radio signal event. Largely, such advances are the result of the more rigorous signal processing made possible by Moore's Law increases in the utility of semiconductor parts.

Parallel developments are taking place (or have taken place) in other radio systems, such as IEEE standards for 802.11n Wi-Fi, 802.16 Wi-Max, and the 802.20 Mobile Broadband Wireless Access (MBWA) system.

The advances in efficiency that allow wireless networks to carry more data don't move as fast as they do for wireline networks, so there's no substitute for more spectrum when it comes to making wireless networks faster and more capable.

The advances in efficiency that allow wireless networks to carry more data don't move as fast as they do for wireline networks, so there's no substitute for spectrum when it comes to making wireless networks faster and more capable.

The principal reason that LTE offers greater opportunity to increase data rates is its use of 10 MHz channels instead of the 5 MHz channels used by previous mobile technologies such as EDGE and HSPA. Similarly, the DOCSIS 3.0 standard used on advanced cable Internet systems increases capacity from 40 to 160 Mb/s by "bonding" or combining four cable channels together, and advanced VDSL+ systems sometimes employ copper "pair bonding" to double capacity. The greater the amount of spectrum a wireless system can use, the higher the data rates it can deliver.

3.4. *Manageability*

Mobile network management is a topic that includes aspects that overlap with Internet management, such as the interconnection of separately owned and operated mobile networks with each other, but it also covers areas that are unique to physical networks. One of the areas unique to physical networks is something we'll call micro-contention. This problem can be illustrated by taking a simple example.

On a Wi-Fi network, the Wi-Fi Access Point (the function that handles the Wi-Fi part of a wireless router) and each station (laptop or smart phone) transmit and receive on the same radio frequency. Stations have knowledge of their own backlog of packets awaiting transmission, but no knowledge of the backlog in other stations. A station or Access Point that wishes to transmit a packet listens to the radio frequency channel before transmitting. If this listening – called "carrier sensing" – shows that a packet is on the air, the station waits for the period of "dead air" that ensues at the end of the packet.

Mini-Tutorial: Why is the wireless downstream faster than the upstream?

Mobile systems typically use separate frequencies for downstream (to the handset) and upstream (to the tower.) Since the tower is the only user of the downstream channel, it can use all of it without any overhead. The upstream is a shared resource that must be allocated either by scheduling or by contention. Scheduling systems can divide the upstream channel a number of different ways, by frequencies, sub-carriers, time slots, or codes, but every scheme limits each device to less than 100% of the channel's capacity. The DOCSIS cable modem system works like a mobile network in this regard.

Having found a quiet period, the Wi-Fi device then waits for an additional period of random duration before transmitting. The random delay is intended to prevent two or more stations what were waiting for the end of the same packet from beginning to transmit at the same time. If the channel is still quiet at the end of the delay period, the station begins transmitting. The random delays range from 0-31 multiples of the "slot time", a unit representing the maximum latency from the network edge to the Access Point, generally 20 microseconds. If two stations begin transmitting at the same time, their packets will corrupt each other, in much the same way that two people speaking at the same time confuse human listeners; this is called a collision.

The stations will not know that this has happened immediately, because they can't receive and transmit at the same time for technical reasons.[42] They will discover the collision eventually, because the Wi-Fi protocol requires the receiving station to acknowledge its successful receipt of a packet by sending a short acknowledgement packet.

If the sender doesn't get an acknowledgement of its packet in the expected time, it will re-randomize for a longer period and then retransmit. This sequence of events resolves an issue that engineers call "the multiple access problem". This particular method is reasonably effective for Wi-Fi because its very limited transmit power enables the slot time to be very small; in larger networks, this method becomes inefficient because the probability of collisions increases with the network's range and with the size of the user population connected to a given Access Point. (As Wi-Fi moves to higher speeds, carrier-sensing overhead is becoming problematic, however.)

Mobile networks have tended to employ altogether different strategies for resolving micro-contention, and these methods utilize characteristics of phone calls. Unlike web access, which occurs in short bursts of activity separated by longer periods of inactivity, digitized phone calls consist of a series of packets separated by regular intervals; this kind of network access is called "isochronous" in contrast to "asynchronous" random access. Knowing that a mobile phone is engaged in a call allows a bandwidth manager in the mobile system to reserve bandwidth on these regular intervals, which does away with the inefficiency of collisions; an application-aware bandwidth manager knows when the next voice packet is coming, and can reserve bandwidth for it. The Wi-Fi system is 50% efficient at 802.11g speeds and much less at 802.11n speeds[43], but a scheduled system can approach 100% efficiency, even over much larger distances and with greater populations of active users.[44] Scheduling systems convert the "dead air" created by the inefficiency of contention into usable bandwidth.

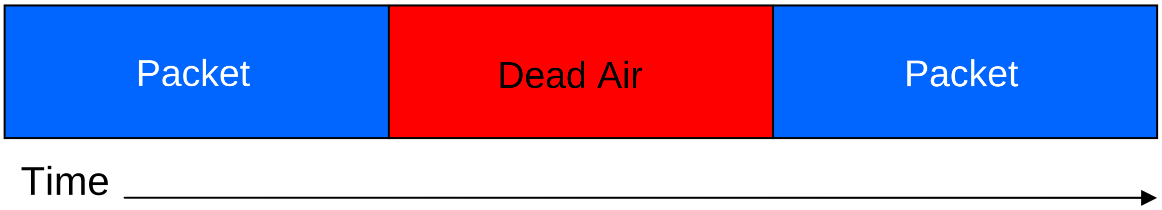

Figure 4: Carrier-sensing overhead in an IEEE 802.11g Contention System

The reason the 802.11g system is 50% efficient is that the dead air time is roughly equal to the time needed to transmit the maximum packet at the network's maximum speed. At higher speeds – such as those enabled by the new 802.11n standard – the dead air time remains the same, but the time required to transmit the maximum packet declines quite dramatically. The top rate of 802.11g is 54 Mbps, but the top rate of 802.11n is (theoretically) 450 Mbps. With no other changes, the dead air overhead would consume nine-tenths of network capacity at the highest 802.11n rate.

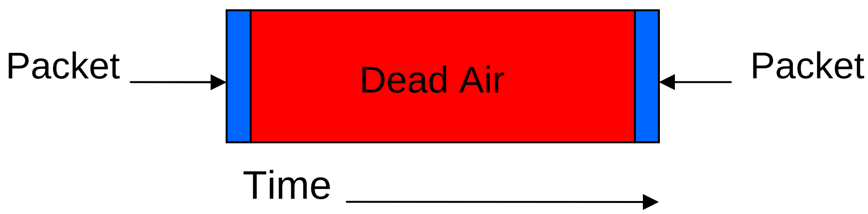

Figure 5: Carrier-sensing overhead in an IEEE 802.11n Contention System

The arithmetic of contention strongly suggests that higher-speed wireless networks will not be the edge-managed systems envisioned by some Open Spectrum advocates. High-speed multiple access wireless networks will be managed by bandwidth managers, as mobile networks are managed today.

Scheduling can help web-based applications as well as phone calls, but this takes more work inside the network. Wireless schedulers typically set aside a recurring period for random access in order to allow users to access the Internet. Collisions can occur during the random access period, but they don't have any impact on telephone callers using scheduled intervals. The mobile scheduler can increase the efficiency of Internet access – eliminate collisions for everyone's benefit – if it knows the network requirements of the particular Internet application. Because the Internet assumes that its constituent physical networks lack a scheduling feature, however, the core Internet Protocol doesn't know how to communicate scheduling requirements from an application that requires them to a network that can easily provide them. The Type of Service bits in the IP header (shown in

Figure 1) don't have a setting for "periodic access." An all-IP network such as LTE has to resolve this problem.

Several types of Internet applications would be helped by bandwidth scheduling, in particular VoIP and audio-video streaming. Any application that deals with sound has to process inbound packets with little delay. While some audio-video applications can buffer large numbers of packets in order to compensate for network delays – YouTube and Netflix Watch Instantly do this – others such as TV channel surfing, video conferencing, and VoIP don't benefit from buffering. In fact, Internet-based VoIP applications such as Vonage and Skype have the same scheduling needs on wireless networks as network operator-managed cell phone calls. As network speeds increase, as they must in order to provide satisfactory web access for mobile platforms, the value of scheduling also increases. Even advanced wireless access methods such as CDMA, beam forming and Spatial-Division Multiple Access (SDMA) need a rudimentary scheduler in order to eliminate the collisions caused by multiple use of shared facilities such as CDMA codes, radio frequencies, and OFDM sub-carriers.

Given that the IP doesn't have the semantics to communicate with a scheduler – it lacks the vocabulary with which to describe streams of repeating events – the question of how real-time communications-oriented applications will work on an all-IP network is a serious issue. In order for next-generation wireless networks to reach their full potential for supporting both data-oriented and communication-oriented applications a resolution must be found, and quickly. The absence of this feature, and its implications, has long been understood by Internet analysts. Tim Wu, the law professor who coined the term "network neutrality", addressed it in his seminal paper "Network Neutrality, Broadband Discrimination:"[45]

> Proponents of open access have generally overlooked the fact that, to the extent an open access rule inhibits vertical relationships, it can help maintain the Internet's greatest deviation from network neutrality. That deviation is favoritism of data applications, as a class, over latency-sensitive applications involving voice or video. There is also reason to believe that open access alone can be an insufficient remedy for many of the likely instances of network discrimination.

Regardless of its many benefits, an open access rule does not correct the Internet's intrinsic design bias against latency-sensitive applications running over wireless networks, and the traditional remedy employed on wireline networks – throwing bandwidth at the problem – has little utility in the wireless space. We'll address practical solutions to this dilemma shortly.

3.5. *Innovation*

It's well understood that open access networking regimes spur Internet innovation While this factor is neither an absolute requirement nor a sufficient condition for achieving the optimal level of innovation, it's extremely important. Other key factors are low entry barriers, easy system extensibility, functional richness, and platform reliability.

When we consider the opportunities for innovation on IP-enabled mobile networks, we need to take stock of the extent to which IP is truly open. While it's true that IP enables every user at every endpoint of the Internet to communicate with every other user at every other endpoint, it is constraining with respect to the parameters of their communication. IP was engineered around the common capabilities of a limited number of networks in an era in which semiconductor chips were much more expensive and limited than they are today. The networks were very limited by current standards. PRNET, for example, supported basic mobility within the extent of PRNET, but not telephony.

IP's original Type of Service parameters didn't support any network capabilities that didn't exist in the four target networks at the time IP was designed, for obvious reasons. IP's "future proofness" consisted primarily of a means by which the first version of IP could be replaced wholesale with a newer version. This is an extremely difficult process, however, as we've learned from the decade-long effort to replace IPv4 with a larger address cousin, IPv6. Network architect Steve Deering has described the Internet's innovation environment as an hourglass, in which a diverse set of applications run across a diverse set of networks, all constrained by the narrow waist of IP.[46]

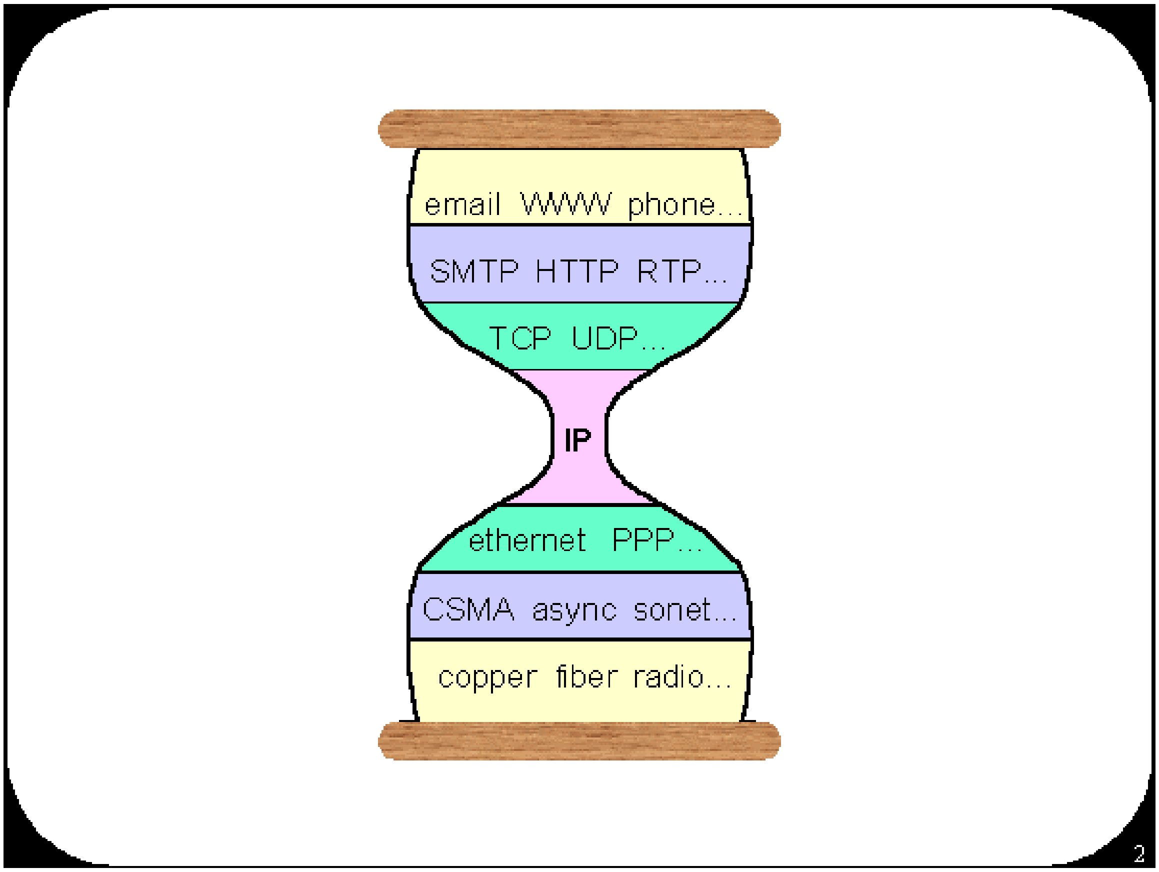

Figure 6: Deering's Hourglass

As Deering pointed out in his 2001 presentation to the IETF, one of the functions of IP is to “virtualize” the physical networks that carry IP packets. Virtualization hides the differences between the networks, but in so doing forecloses the ability of some applications to make efficient use of some network features. The uneven range of capabilities in physical networks makes this outcome unavoidable: IP is a physical network interface, and when physical networks have diverse capabilities, it must choose between uniformity and capability. Uniformity is a maximum hiding of differences, and capability is a minimum hiding.

The clear bias in the Internet engineering community is to err on the side of uniformity. This bias compromises network efficiency in the interest of network generality. This bias was productive for the Internet of the past because the shared resource, bandwidth, has grown steadily cheaper. In principle, wireline bandwidth is only limited by the number of silicon atoms that can be transformed into glass fiber and semiconductor chips, and the willingness of network users to pay for them.

Internet uniformity adheres to a least common denominator assumption about network features, but it can just as easily be conceived the other way around. Mobile networks don’t rely on packet schedulers because they’re more capable than wireline networks, but because they're more constrained. When we can't eliminate network congestion by adding bandwidth, we need to schedule packets in order to boost the overall efficiency and effectiveness of the wireless network up to accepted wireline standards. Consequently, VoIP and similar applications only need to communicate scheduling desires to the wireless networks they traverse, not to every network in its end-to-end path. The Internet allows this kind of communication through specialized QoS protocols such as Integrated Services[47] and RSVP.[48] In the next section, we'll describe how these protocols may be employed to support voice over LTE.

3.6. *Specific Example: Voice over LTE*

The 3rd Generation Partnership Project (3GPP) is designing LTE. While the design has been heavily influenced by the needs of voice at every level of the architecture, the final details about how to support it over IP are not yet determined. For the time being, LTE demonstration projects will support voice by falling back to the pre-LTE mobile voice standards. Since LTE is meant to be a smooth upgrade, this is not burdensome for the time being. In the long term, however, a method must be developed for voice over LTE. The official plan for LTE calls for a system known as IP Multimedia Subsystem (IMS) to provide voice support, but the IMS standard is not yet complete and equipment to support it is very expensive. The consensus position within 3GPP is that an interim system will transition to IMS; the two candidates are One Voice[49] and Voice over LTE via Generic Access (VoLGA).[50] Essentially, VoLGA is a short step toward IMS and One Voice is a longer step. One Voice, recently renamed Voice over LTE or VoLTE, has all the momentum.[51]

IMS is a work in progress as it seeks to accomplish two contradictory goals: on the one hand, it wants to use low-cost, general-purpose hardware developed for classical IP networks, and on the other is wants to provide a functionally rich platform for deploying and monetizing media services. Hence there's been great deal of tussle from one release of IMS to the next regarding function placement: rich services can be provided within the IMS network, and low-cost services can be deployed outside it.

IMS access to specialized network QoS capabilities is provided by a function called the Resource and Admission Control Subsystem (RACS). As part of the processing of call requests, IMS reserves bandwidth through RACS, a function that bypasses IP and communicates directly with the radio network's scheduler and the IMS Service-based Policy Decision Function (SPDF). SPDF is connected to the IETF-standard Session Initiation Protocol (SIP), so we can understand IMS as supplementing the standard Internet protocols with application- and network-specific interactions.

Once bandwidth is reserved, an RACS-aware function in the wireless device responds to scheduling with the appropriate voice packets. It's in the interests of the wireless device to use scheduled bandwidth appropriately; not only is it a scarce resource with higher cost-per-byte than unscheduled bandwidth, it's also managed in smaller increments that wouldn't serve the needs of applications such as web surfing very well. Scheduled bandwidth is generally subject to bargaining between application and network: the network permits low-delay access to the network, but only for short periods and only on an extended schedule. This is similar to the ability of an administrative assistant to interrupt a busy executive to deliver short messages such as “you had an urgent call from your broker”. This sort of access wouldn’t be appropriate if the administer wants to have an extended conversation about buying a car while the executive is meeting with a client.

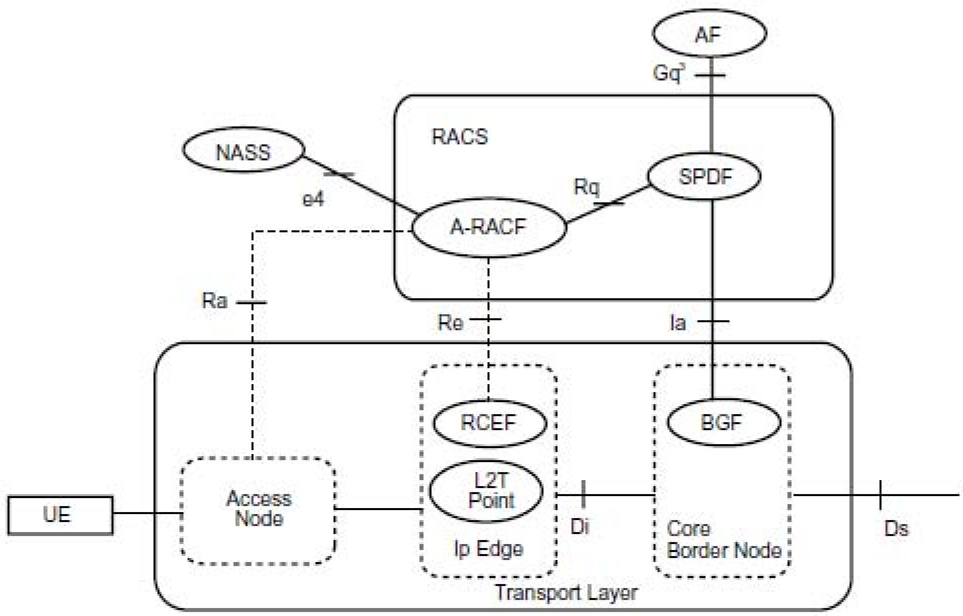

Figure 7: Internal Architecture of the RACS model[52]

Whether the IMS model is consistent with Internet architecture is a complex question. Some advocates insist that the Internet architecture forbids application-specific features in the network; if such features can’t be placed in the (IP) network, they can only be implemented in specialized features of physical networks such as IMS. By that understanding, IMS is consistent with Internet architecture, even though it’s clearly not consistent with Internet tradition.

An alternative approach to providing voice over LTE would use IETF standards such as Mobile IPv6 along with RSVP and Integrated Services. Even with these protocols serving as connectors, mobile networks would still need to supply their own servers to complete the roaming, authentication, and accounting functions necessary for network operation. In the same way, fully-IETF content services need to bring their own content and billing to systems that use HTTP to connect to consumers. A more standards-compliant solution may win out in the long run because it would lower network operation costs, but it’s not available at present.

3.7. *Status of Internet Design and Wireless Networks*

Internet standards don't fully support the range of capabilities required by mobile users and wireless network operators, so the wireless engineering community is at work on the necessary supplements. This process of supplementation and substitution isn't unique to wireless networks: new wireline applications such as BitTorrent and Skype also substitute idiosyncratic protocols for Internet standards. In some cases, such innovators seek IETF approval for their protocols, as BitTorrent is currently doing with their uTP protocol; Skype is not going in this direction, however.

The Skype system uses a P2P call setup procedure based on the KaZaA system created by Skype's founders before the formation of Skype. The KaZaA method penetrates firewalls better than the IETF method, SIP. Skype also uses proprietary methods of encoding phone calls in digital formats, and relies in part on the license fees their proprietary methods generate to provide cash flow; the bulk of Skype's revenue comes from PSTN termination fees, however.[53]

As is the case for Skype and P2P, mobile network innovation takes place beyond the Internet's design horizon. Wireless operators hope that IP technology will ultimately catch up and fully assimilate the mobile capability, if for no other reason than to lower the costs of wireless network equipment. The dream of lower-cost equipment was the prime motivator for the all-IP LTE design.

The design of the Internet presumes that the underlying physical network has a set of capabilities that closely approximate those of the classical Ethernet local area network in the 1970s. Classical Ethernet was a stationary network, with a single service level, high speed, low intrinsic error rate, and variable delay; it was the prototypical "best-effort" network. The Ethernet assumption doesn't adequately capture the nature of today's physical networks.

Wireless networks, as we've seen, have the ability to reserve and prioritize the allocation of network bandwidth; they're also mobile and highly variable with respect to delay and packet loss. Ethernet itself has evolved from a shared cable, "stupid" network controlled by end points alone into a system with an active switch at its center which can mediate access requests from all connected stations and assign bandwidth to high-priority uses ahead of low-priority ones according to the seven levels of priority defined by the IEEE 802.1D standard.[54] In today's Ethernet, "best effort" is the second to lowest priority. The IEEE 802.11e Wi-Fi standard defines four levels of priority as well as a reservation-oriented service.[55]

It's no longer necessary or productive to think of the Internet as a best-effort network that operates on a purely end-to-end basis. As David Clark, author of the papers on the end-to-end arguments and former Chief Architect of the Internet, recently said:[56]

> For many users, the end-to-end principle in its literal form is a pain - it means they have to install software and manage upgrades on a PC that is complex and insecure. Much better to take advantages of services that are professionally run…I think the crucial question is not where a function is located (at the end-point or from a service provider somewhere on the network), but the extent to which the end-user will preserve the right to choose providers that they decide to trust. The real question is about trust, not location.

When network users have meaningful choice among providers of network services, many of the most vexing regulatory questions answer themselves.

4. The Mobile Network Infrastructure

Up to this point, we've discussed the mobile network in terms of its protocols and system architecture elements. In this section we'll examine the equipment that makes it operate. The purpose of this section is to develop an understanding of some of the practical limits on mobile networks in order to uncover barriers to network growth that can be alleviated by smart policy choices.

4.1. *Essential Parts of the Mobile Network*

The mobile network consists of three basic parts:

A) A battery-powered mobile handset with one or more radios, each with its own antenna;
B) A base station with one or more radios, each with its own antenna, that communicates with all the mobile handsets joined to its network in an area ranging from a city block to a few miles and bounded by the areas served by neighboring base stations;
C) A backhaul system that connects base stations with the telephone network and then to the Internet core.

Each of these elements shares certain common characteristics that allow them to function as a coherent whole. The **handset** identifies the subscriber, communicating his or her location (among other things.) This function is unique to the mobile network; wired networks identify subscriber by the wire to which they're attached. Because the handset is battery-powered, the mobile network system is designed to conserve handset power. Handsets turn their radios on and off several times a second at precise intervals to sense signals from the base station without consuming the power required for always-on operation. Handsets create the illusion of an always-on connection without actually having one.

Base stations (AKA, towers) were originally built to use a single radio and antenna, but have been redesigned to employ a technique known as "sectorization" that divides the coverage area into three or more sectors (imagine a pizza cut into three equal slices.)

Each sector is served by its own radio and antenna pair (or more than one for MIMO systems.) Sectorization is a technique that increases the effective capacity of a base station, but every sectorization system creates areas of coverage overlap where signals from adjacent sectors overlap each other; handsets are engineered to deal with these situations.

Backhaul is a system of copper wire, fiber optic cable, microwave, or millimeter wave radio that connects base stations to the network operator's wired network. Once the base stations are connected to the operator's wired network, it becomes possible for the operator to distinguish phone calls that are to be routed to the PSTN from Internet packets that are routed to Internet Exchange Points. Backhaul is a particularly serious deployment problem that often does overlooked in the policy discussion around next-generation mobile networks because it's seen as secondary to spectrum. As the data rates of handheld devices increase due to better signal processing and more spectrum, the load on backhaul also increases.

4.2. *Spectrum*

Radio spectrum serves two purposes in the mobile network: it's the indispensible means by which handsets communicate with base stations, and one very common means of providing backhaul. The frequencies for handset coverage are lower than those used for backhaul; lower radio frequencies naturally tend to propagate more uniformly, while higher ones tend to propagate more narrowly (these characteristics can be modified somewhat by antenna design, but not completely erased.) When the goal is to blanket an area with radio waves, frequencies below 4 GHz do the trick. When the goal is to connect a base station to one and only one network concentration point a mile or more away, microwave frequencies above 10 GHz are appropriate. When the goal is to connect a base station to a nearby network concentration point without a license, 60 GHz millimeter wave does the job. The 700 MHz frequencies freed up by the DTV conversion are usable for mobile networking, but are somewhat less effective for data services than frequencies above 1500 MHz; at 700 MHz, the beam is too resistant to reflection for advanced techniques such as Multiple Input-Multiple Output (MIMO) radios to generate optimal efficiencies. Mobile operators can use 700 MHz, but also need additional spectrum above 1500 MHz.

Licensing considerations sometimes motivate network operators to use frequencies for particular applications that may be less than ideal, or to use them in novel ways. An experimental network is building out in Claudville, Virginia using Wi-Fi for the last hundred feet and license-exempt 700 MHz White Spaces for backhaul.[57] The network apparently obtained a waiver from the FCC allowing it to boost White Spaces power above the ordinary limit, which allows a single White Spaces signal to reach an entire group of Wi-Fi access points.

4.2.1. Licenses

While licenses are seen as a nuisance by rural network operators, in densely populated areas they have the advantage of guarding against "tragedy of the commons" effects frequently encountered by Wi-Fi users in high-density housing and conference settings. Every radio frequency channel has a finite capacity in every geographic area, and licenses are one means of limiting access to the number of users and amount of use that the channel can support. Licenses are typically limited by frequency, bandwidth, power, and geographic area in the U. S., and by technology as well in Europe.

4.2.2. Efficient Use

While licenses help to ensure efficient use of radio channels, they're neither necessary nor sufficient constraints. Multiple use of radio channels can be achieved by voluntary adoption of standards, but as we discussed in the previous section on wireless manageability, existing standards aren't as comprehensive as they would have to be in order for high data rate users to share radio channels over distances greater than two or three hundred feet in densely-populated areas. The necessary work to develop open standards for wireless channel sharing with reservations simply hasn't been done.

Even if such a standard existed, investment isn't likely to occur without the assurance that the investor will have reasonably predictable and successful use of radio channels. Licensing has an advantage, at this stage of technical development, for assuring the efficient and deterministic use of radio channels.

4.2.3. Roaming

Roaming makes capacity planning for mobile networks tricky. In principle, any estimate of the number of users within the coverage area of a given base station is subject to error; a "flash mob" may converge in a public space and overload the base station. In practice, these events are rare, so wireless operators have developed formulas for determining capacity while taking roaming into account.

Europe has extensive regulations on roaming that cover capacity and permission, and is considering adopting roaming regulations for mobile data plans in order to reduce roaming charges. The interplay between such requirements and capacity planning is very complex, hence GSMA (the trade organization of 900+ mobile network operators and equipment vendors) opposes data roaming regulations.

4.2.4. Coordination

Such issues as base station density and location and roaming for both voice and data raise the issue of coordination of radio channels across base stations. Wireless signals don't propagate uniformly; they're distorted by physical obstacles, reflected by surfaces of various kinds, and disperse with the square of distance and frequency due to free space path loss.[58] In areas where signals overlap, such as between neighboring base stations or between sectors on a single base station, coordination strategies are employed to mitigate interference. The simplest strategy is to use different radio frequencies on adjoining sectors, just as AM and FM radio stations transmit in different frequencies; this is a traditional analog radio technique.

Digital radio systems such as those used by the mobile network also mitigate interference using different codes on a common frequency. The Code Division Multiple Access (CDMA) networks employed by Verizon and Sprint use a technique developed by Qualcomm, and the 3GPP system used by T-Mobile, AT&T, and the European operators use WCDMA, a licensed variant of the Qualcomm method. CDMA uses digital coding signatures to train radio receivers to the correct transmitter at any given time.

An additional technique is Time Division Multiplexing (TDM), where systems take turns accessing the radio channel per a schedule. OFDM Multiple Access (OFDMA) and Synchronous CDMA (S-CDMA) are additional tools that may be used in the pursuit of coordination, as is the TD-SCDMA system preferred by China, a hybrid of TDM and S-CDMA that allows active management of upstream and downstream bandwidth allocations.

The coordination problem hasn't been completely resolved, but common management of adjacent base stations and the frequency licensing makes it tolerable. The cooperation of the various elements of the system is assured by common ownership.

4.3. *Base Station Siting*

Adding base stations is one solution to mobile network capacity exhaustion, but it's fraught with difficulties. High base station density aggravates the coordination problem, as the greater the density, the greater the likelihood of signal overlap and interference. The techniques described previously have operational limits in terms of the number of codes, frequencies, and time slots available for assignment. Adding base stations also requires backhaul connections to more places, as well as local zoning permission. Zoning raises NIMBY considerations, especially in areas in which citizens are extremely sensitive to aesthetics. For this reason, tower designs have been developed that blend into background features.[59] But even with this creativity on the part of tower providers, many jurisdictions still arbitrarily limit new towers.

Figure 8: Cell tower disguised in cross by Larson Camouflage.

The FCC's recent adoption of "shot clocks" for local zoning decisions is extremely helpful in this regard, and should help clear the backlog of tower siting requests that currently exceeds 760 according to the CTIA.[60] Nevertheless, the shot clock only requires a faster decision, and even with this improvement many localities still put up barriers to cell tower siting. Moreover, each new base station requires backhaul, which has its own problems.

4.4. *Backhaul*

Before we define and address the options for mobile network backhaul, it's useful to take a step back and review how the Internet is put together. In common discourse, the Internet is a single, coherent network that exists in its own space. ISP networks are supposed to "access" this Internet, so their terms of use are believed by some to be regulatable apart from the terms under which the Internet operates in its far-away place. In fact, this is a misleading characterization.

The Internet does not exist apart from ISP networks, as it's not so much a thing or a place as a set of agreements that enable separately owned and operated networks to exchange information under mutually agreeable terms. When subscribers purchase Internet services, they're not *accessing* this system as much as *joining* it. The claim that regulations affect access but not the Internet system as a whole is a distinction without a difference. Network neutrality advocates know that the Internet is an end-to-end system, so there is no part of the path between one Internet-connected system and another that is not part of the Internet. A regulation on one part of the Internet is a regulation on the Internet as a whole.

Whenever a new physical network joins the Internet, the virtual network – the Internet itself – expands and assimilates the new physical network. Terms of interconnection therefore must reflect the characteristics of physical networks and levels of service their owners, users, and applications require.

It's more productive to think of the Internet as an ecosystem than as some sort of advanced telecom network. The following will examine the ways that different aspects of the Internet ecosystem interact within the medium of mobile networking.

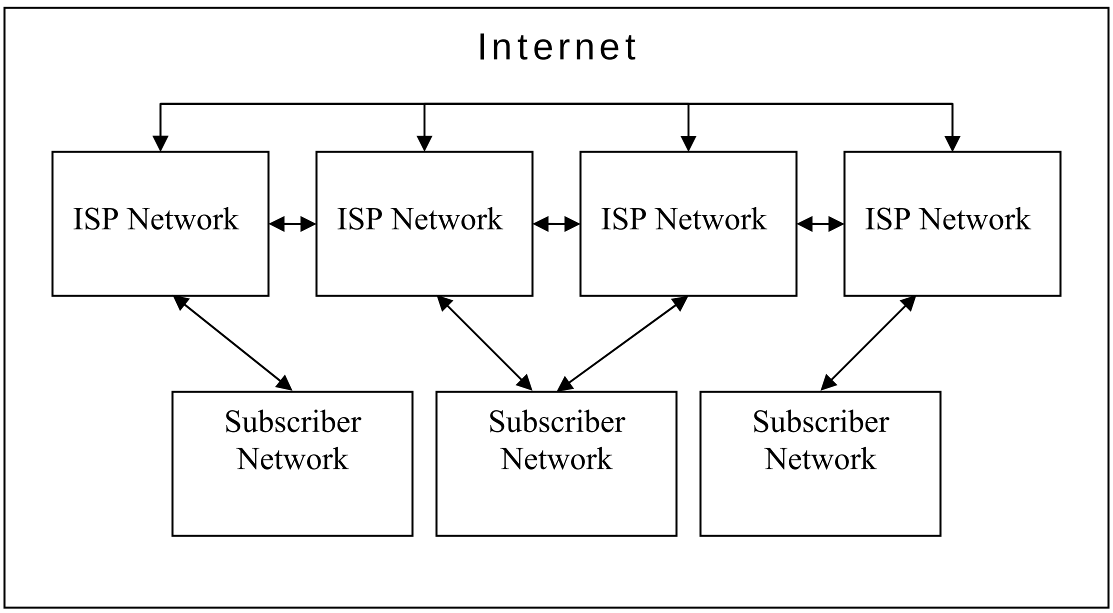

Figure 9: Internet is a Virtual Network composed of Many Physical Networks.

Consequently, one of the crucial aspects of the mobile infrastructure is the element that joins mobile networks to the Internet; this part of the mobile network is called the "backhaul". The backhaul combines or "aggregates" network traffic to and from towers into a single point of concentration under the network operator's control, and then connects to Internet Exchange Points (IXPs) where IP packets are exchanged with other networks.

4.4.1. Technologies

The most valuable backhaul technology is optical fiber; it permits a higher bandwidth horizon, lower latency, and greater resilience than any of the alternatives. Once fiber optic cabling is installed, it pays immediate dividends in every other aspect of network operation: power consumption is lower, reliability is higher, and termination equipment is cheaper. However, installing fiber can be expensive: costs vary from $5,000 to $500,000 per mile depending on the development state of the installation area (rural is cheaper than urban), aerial deployment vs. trenching, trenching requirements (under pavement vs. under dirt), and access to ducts and conduits. When these costs are prohibitive, network operators turn to alternatives such as copper pairs, microwave, and millimeter wave.

Telephone company twisted pair has been the historical mainstay for mobile tower backhaul. The copper-based T1 standard for digital telephony is ubiquitous and mobile is still a telephony-oriented system. T1 has capacity issues – its data rate is only 1.44 Mbps – but these can be overcome by bonding multiple pairs. Eight T1s is 11.5 Mbps, sufficient for mobile technologies up to HSPA. When copper is not sufficient and fiber is too expensive to install (both are often the case in crowded city centers,) the next alternative is microwave.

Microwave frequencies are licensed, so the network operator who invests in a microwave backhaul system doesn't have to worry about the system's capacity. As long as all nearby microwave users are operating within the terms of their licenses, capacity will be as expected. Microwave is a line-of-sight technology that performs in a manner very similar to cables, connecting one pair of communicators exclusively, without spillover or interference to or from others.

Despite the economy of microwave signals, in come urban areas all microwave frequencies are already in use. A new entrant – or an old entrant with a new tower – can sometimes persuade an incumbent to alter his signal path in a dogleg pattern to create an interference-free path by installing an intermediate transmitter and receiver.[61] This sort of negotiation is common in high-density settings.

Where all other options are exhausted, millimeter wave systems are available at 60 GHz and above in both license-exempt and lightly licensed regimes. These systems consume a great deal of power for the distances they cover – the higher the frequency, the greater the power to cover a given distance – but they're so rare and have such highly focused signals that interference with other millimeter wave systems isn't an issue.

4.4.2. Specific Examples: GigaBeam and Ceragon

Microwave backhaul equipment is available at prices ranging from $10,000 to $100,000 depending on speed, distance, and power level. Transmission systems capable of achieving the 100 Mbps speeds needed for 4G backhaul cost roughly $1,000 - $2,000 for each mile of range from 6 to 30 miles, plus a base cost that increases with frequency.

Ceragon Networks is an Israeli microwave technology producer offering a wide range of wireless backhaul and distribution systems. Unlike millimeter wave competitor GigaBeam,[62] Ceragon is a profitable company, although 2009 was a hard year for them. Their product line is a mix-and-match set of modular components consisting of RF air interfaces for the 6-38 GHz range, protocol processors supporting Carrier Ethernet, Multi-Service, and TDM, and management systems. Their flagship product, FibeAir IP-10, is a carrier-grade wireless Ethernet providing “migration to IP with the highest possible capacities at the lowest overall cost.”[63]

Figure 10: Cergagon FibeAir® IP-10 Switch

Ceragon systems are designed like traditional Ethernet switches with “optional stackable radios” and external network management capabilities. They’re meant to provide mobile operators with a platform for migrating traditional backhaul networks to IP. Systems range in speed from 100 Mbps to 1.6 Gbps and in distance from 6-30 miles.

Millimeter wave systems produced by GigaBeam and others show some promise in a limited set of backhaul applications, but they’re not without challenges. Distance is a problem, as these systems are typically limited by power requirements to 1000 meters on the license-exempt 60 GHz spectrum. Their principal use is currently connecting two adjacent buildings when it’s impractical to install or lease optical fiber, but GigaBeam systems are also in use in Google’s public Wi-Fi network in Mountain View, California, where they provide backhaul.

Figure 11: GigaBeam installation providing backhaul for Google's Mountain View Wi-Fi Network

4.4.3. Backhaul Bottom Line

Termination equipment costs and power requirements are much higher for wireless than for fiber, but wireless installations are faster, cheaper, and easier. It's hard to argue with the long-term economic benefits of user-owned fiber networks, but ease of installation can tilt the scales in favor of wireless systems that enable the operator to reach paying customers sooner even if operational costs are higher. One Velocity, a GigaBeam customer in Las Vegas, Nevada, was able to design and install a small Metro Ethernet and win customers in four months with a lightly licensed 70 GHz system, for example.[64]

Fiber to the tower, however, has ancillary benefits for fiber rollout generally that shouldn't be ignored even though they're hard to quantify at present. It's rare for a fiber rollout not to include more fibers than the immediate application requires, and these can generally be made available on some basis to other applications such as fiber to the home or office.

5. Emerging Mobile Internet Applications

First-generation mobile Internet applications were limited to e-mail and surfing portions of the web acceptable to the Wireless Application Protocol (WAP) system. WAP 1.0 didn't connect to the web directly, but to a gateway that reformatted web pages for the small screens of the wireless "feature phones" that preceded today's smartphones. WAP 2.0, released in 2002, provides for end-to-end web access without the gateway but still allows the operator to add subscriber information to web transactions. Leading edge smartphones such as the iPhone, the Palm Pre, the Androids, and Blackberries access web sites in much the same way that laptop computers do, with the additional benefit of mobility. Hence, they're already a richer platform for applications than desktop computers: they reach the whole Internet, the Global Positioning System (GPS), the telephone network, and in some instances the TV network as well. Consequently, these smartphones are the platform on which the most interesting and useful applications are emerging.

5.1. *Content Delivery*

At first impression, content applications seem to work the same way on mobile networks that they do on wireline networks: a user locates the content, clears any licensing hurdles that may be associated with it, starts a transaction to transfer the content to his or her wireless device, waits a while, and finally enjoys the content. If the content is on a public web site and the device is a laptop, the user deals with network access and content licensing in the same way these things are done on the wireline Internet, by paying a fee to an ISP and perhaps a second fee to the content supplier.

5.1.1. Example: Kindle

This is not the only business or technical model, however. For example, if the content is a book on Amazon and the device is a Kindle, a different procedure takes place: The user pays a fee to Amazon, and Amazon takes care of the wireless transfer using one of a number of networks with which it has agreements. The transfer takes place transparently, without any need for the user to have a financial arrangement with a network operator. This alternate business model is possible because wireless networks are more ubiquitous than wireline networks. Instead of connecting to a network through a dedicated cable connecting the ISP to the user's home or office, the user can connect, in principle, to a number of wireless networks through a single antenna.

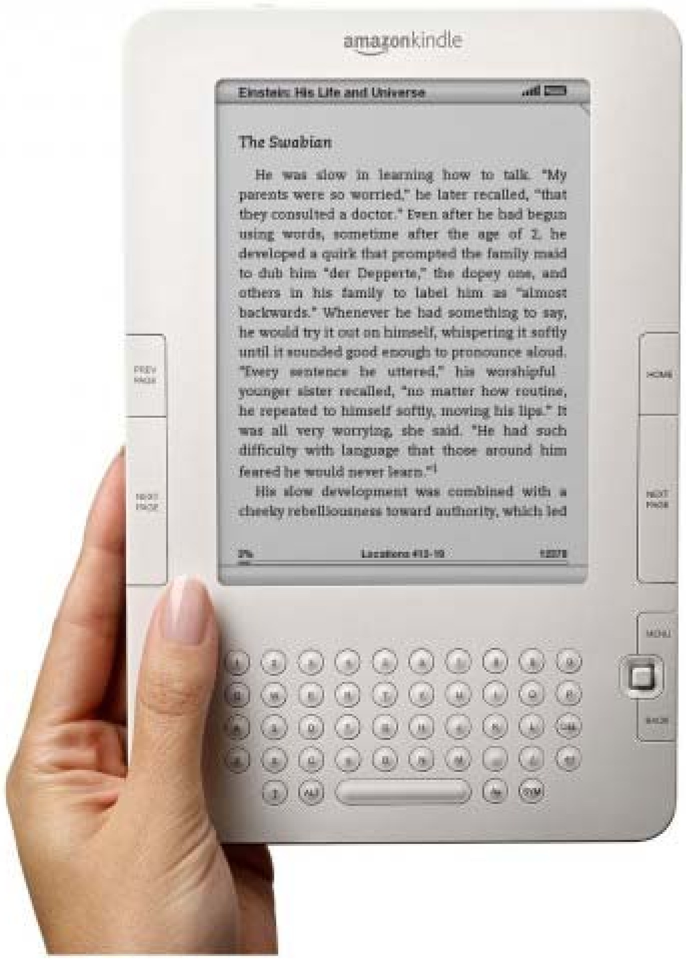

Figure 12: Amazon Kindle

5.1.2. Benefits of Ubiquity

The ubiquity of wireless signals enables a new range of business models for content delivery that simply don't exist in the wireline world. These business models enable users to enjoy benefits of the purchasing power of large content suppliers, in much the same way that users enjoy similar benefits when buying tangible goods from large on-line retailers, and they enjoy further economic benefits from customized transport services. In comparison to wireless telephony, e-book delivery is not an urgent service; the individual data packets that carry the book don't need to arrive in less than 150 milliseconds, so the wireless network can deliver them during periods when the network would otherwise be idle. Advocates of the anti-discrimination rule for wireline ISPs say the rationale is to prevent potential abuse of market power by the holder of an effective monopoly over access to a particular user. This fear doesn't fit the wireless scenario. In fact, the ability to employ economically and technologically efficient price discrimination opens up whole new business models that benefit the consumer.

The ubiquity of wireless signals enables a new range of business models for content delivery that simply don't exist in the wireline world.

Even when a user transfers content from a web site to a laptop, mobile wireless networking lowers switching costs, enabling users to choose from a number of wireless providers without signing a contract or making a heavy investment in equipment. USB 3G modem prices start at less than $100, and a number of operators offer pay-as-you go plans in increments of $10.[65] People who have smartphones with data access can often tether their laptops, and laptop users can access free or paid Wi-Fi networks at public hotspots, some of them aboard commercial airlines. Clearly, the ubiquity of wireless networks has given rise to experimentation with business models by network operators and content suppliers in ways that the traditional wireline Internet hasn't.

5.1.3. Innovation from Differentiation

In addition to ubiquity, it's important to stress that the nature of wireless packet protocols has itself helped spur wireless innovation. If a blunt packet non-discrimination mandate were in effect, wireless operators would not be able to sell "bulk mail" packet delivery to firms like Amazon at attractive prices. Even though the Kindle application doesn't need low-latency delivery of each and every packet, network operators forced to comply with a non-discrimination mandate would be required to offer the lowest latency QoS to every application, even those that don't need it. This is the goal of David Isenberg, a net neutrality advocate who has written that it's the ideal toward which networks should trend:[66]

> But suppose technology improves so much that the worst QOS is perfectly fine for all kinds of traffic, without a repertoire of different data handling techniques. Suppose, for example, that everyday normal latency becomes low enough to support voice telephony, while at the same time allowing enough capacity for video, plus data integrity strong enough for financial transactions. This would be a true Stupid Network—one treatment for all kinds of traffic.

This may be a lovely ideal for network engineers dreaming about abstract networks, but it's neither practical for real networks nor a valid foundation for legally enforceable regulation in the year 2010. For one thing, there's no demonstrable benefit in designing the intelligence out of our networks. The assumption that networks can be smart or fast but not both is a holdover from earlier technology battles in which engineers had to make choices between network capacity and control. In the old days, semiconductor processes were so limited that it was literally impossible to fit the logic needed for speed on the same chip with the logic needed for management.

This is no longer the case in the wireless domain where network speed lags behind digital processing capacity. Networks are combinations of digital and analog technologies, and we can process information by purely digital means much faster than we can communicate it over the air. The Isenberg dichotomy between smart, slow networks and faster dumb networks is no longer relevant, if it ever was.

5.2. *Wireless VoIP*

Wireless VoIP – Skype and similar systems – has been at the forefront of much of controversy around regulating the wireless Internet. It's a subversive application in more ways than one. It directly attacks the revenue stream of mobile network operators who built voice-oriented networks on the assumption that per-minute mobile telephony and SMS would provide the revenue to support network operation costs. Adding insult to injury, wireless VoIP is also more resource-intensive than cellular telephony because it uses the less efficient contention-based packet data service instead of the reservation-based circuit service. Wireless VoIP is also more resource-intensive in terms of overall bandwidth consumption because of greater per-packet overhead and a higher top-end capacity to translate bandwidth into audio quality.

Mobile networks and the Internet both possess an engineering feature known as rate adaptation that adjusts applications' appetites for communication resources down to available supply; Jacobson's Algorithm does this for the Internet, and such systems as CDMA do it for mobile. In addition to these systems, conventions exist in both worlds regarding limits and exceptions to rate-adaptive algorithms. In the Internet, real-time communication applications rely on UDP, a protocol that is exempt from Jacobson backoff, and in the mobile world phone calls have a sharp upper limit on the amount of bandwidth they can consume.[67] VoIP thrives on the Internet in the space exempted from end system congestion management, and will ultimately thrive on mobile networks by delivering better call quality in some circumstances and cheaper calls in others. Closer coordination between VoIP and mobile network internals would provide a better overall experience for more users, as well as a revenue stream for VoIP providers that they will otherwise lack when termination fees to the PSTN dry up. Skype currently earns most of its revenue from fees charged to users for connecting to the PSTN, but this revenue stream won't last forever; once we transition from PSTN to VoIP, it will evaporate.

Phone calls made through mobile networks aren't handled the same way as web access. Mobile networks handle calls through a procedure that mirrors the PSTN in all relevant respects, verifying the account and reserving resources end-to-end before ringing the called party. The mobile network also reserves resources for roaming, and over-reserves communication resources to allow for the higher rate of retransmission through the wireless medium. Each of these procedures incurs real costs. The mobile network business model also assumes that callers are willing to pay more per bit than data services users; as calling plan revenue declines, it will be necessary for them to raise data plan prices.

5.2.1. Example: Skype on the 3 Network

Wireless Skype doesn't work the same way that conventional mobile telephone service does. Its call setup procedure is an idiosyncratic system originally developed for KaZaA, the main purpose of which is to circumvent corporate firewalls; it complies with neither IETF nor ITU standards, and isn't even public. Once a call has been established, Skype relies on best-effort packet delivery and its own brand of error recovery to ensure call quality. In many cases, this system produces satisfactory results, and may even yield higher quality calls than standards-based mobile telephony, thanks to the proprietary Skype wideband codec that simply consumes more network bandwidth than a conventional call. Nevertheless, under the most challenging conditions, most likely to occur on mobile networks, the Skype system is unlikely to provide the user with a good calling experience; it has to contend for communication bandwidth on equal terms with web surfing and P2P file transfers and often loses. The struggle over the permissibility of the Skype application on mobile phones hides the fact that many users are likely to find it an unsatisfactory experience.

Figure 13: 3 Skypephone

There is at least one mobile network on which Skype should work well, the 3 network deployed by Hutchison in parts of Europe and Australasia. 3 runs a 3G-only network, offering a phone with Skype pre-installed to its users. 3's service offers free Skype-to-Skype calling within the limits of standard data plans, which are quantity-based.

Reviews of the 3 Skype phone are decidedly mixed. Early adopters were pleased,[68] touting call quality that was often better than PC-to-PC Skype calls, but as the installed base has grown, call quality has apparently declined.[69] Wireless engineering would predict just such results: best-effort obtains low latency on less crowded networks, but not on networks loaded at greater than half of design capacity. In order for Skype users to get the same level of reliability associated with mobile calls, it will be necessary for it to use better-than-best-effort transport under load, which will raise network cost, at least for the wireless portion of the Skype call. The 3 network "cheated" a bit in any case, installing dedicated servers to bypass the Internet backhaul for Skype-to-Skype usage.[70]

5.2.2. Example: Other Ways to Skype

In addition to the way the 3 network supports Skype, two other approaches are used to bring Skype to mobile networks. One approach, currently used on the iPhone, is to allow Skype full access to broadband networking through Wi-Fi and limited access to the mobile network using the customer's voice plan. This approach provides excellent quality of service since Wi-Fi has access to more bandwidth than the mobile network; Wi-Fi channels are 22 MHz wide, while mobile channels are only five to 10 MHz. Wi-Fi also has more limited range than mobile networks, so it shares its greater channel bandwidth among a smaller geographic area and thus to a smaller group of users. When there is no Wi-Fi available, the use of calling plan minutes invokes a higher QoS level in the mobile network, which improves Skype quality under conditions of network congestion.

An additional approach to mobile Skype, just announced by Verizon, simply uses the mobile data plan service to transfer Skype packets. Under conditions of low network load, this approach should provide the Skype user with an experience whose quality exceeds normal cell phone quality, but under high load it will probably be less good. This is the pitfall of any implementation of a delay-sensitive application over a network service that doesn't limit latency.

5.2.3. Economic Implications of Skype

The financial model for mobile networks takes into account the value that users place on different network services and cross-subsidizes in order to keep the costs of basic calling plans as low as possible. As users converge on data plans and stop using calling plan minutes and high-priced SMS services, the overall formula for mobile network services will naturally change. Prices for a basic level of data access are likely to decline, with operators recouping costs and earning profits from high-volume users and users of highly valuable services that require low latency. Over the long term, costs of operating networks will continue to decline on a per-bit basis, but mobile will decline more slowly than wired for a number of technical reasons. Operators will adjust pricing in a number of ways to maintain profitability and competitiveness, not always as economic theory says they should. At the end of the day, revenues will be lost as users shift from calling plans to data plans, and data plan prices will be adjusted accordingly.

5.3. *Wireless Video Streaming*

Wireless video streaming obviously stretches the capacity of 3G mobile networks, but it's an enormously popular application. Top-end smartphones are highly resource intensive.[71] If web- and video-oriented smartphone applications have the ability to share bandwidth efficiently and effectively, they have failed to demonstrate it so far, with the exception of

Slingbox, whose rate-adaptation algorithm has become wireless-friendly.[72] Further advances in network support of video streaming can come about from the increased adoption of multicast models and from allocation of secondary spectrum as we've previously suggested.

The barrier to better support of wireless video streaming is less economic than technical; a number of means are available to address the problem, and they typically represent a tradeoff between less user convenience (multicast only benefits popular programming) and additional spectrum and/or radio towers. This application is likely to generate friction for some time.

Conventionally, Internet-based video programming is delivered over the wireline Internet as on-demand unicast streams from CDNs to subscriber systems such as generic PCs, Home Theater PCs (HTPC), and home entertainment devices such as DVRs, Blu-ray players and TV sets with embedded video streaming capability. Unicast is a form of communication in which each sender/receiver pair has a unique copy of the program stream; it differs from the broadcast or multicast model in which a single program stream serves multiple consumers. The unicast model offers the advantage of personalization in terms of time and content; in principle, a unique set of ads can be incorporated into each unicast stream, although this model is not in common use today. Multicast is much more efficient in terms of communication resources, and is therefore more easily scalable to large audiences. Unicast is interactive, while multicast is generally a one-way system. In an analogy to the wired world, cable TV is multicast, whereas web video streaming on services like Hulu is unicast.

Because video transmission involves orders of magnitude more transmission of bits, Video is a resource problem for bandwidth-limited mobile networks, so new systems have been developed that utilize efficient multicast for popular live programming with a limited amount of personalized content. Qualcomm's MediaFLO and MobiTV's Connected Media Solutions are two examples.

5.3.1. Example Video Streaming System: MediaFLO

MediaFLO is a novel, proprietary system developed by Qualcomm that utilizes the 6 MHz of licensed bandwidth at 716 MHz that was formerly occupied by UHF Channel 55. The FLO consumer interacts with the system through the mobile phone network, and receives FLO programming at 716 MHz regardless of which mobile network he or she uses.

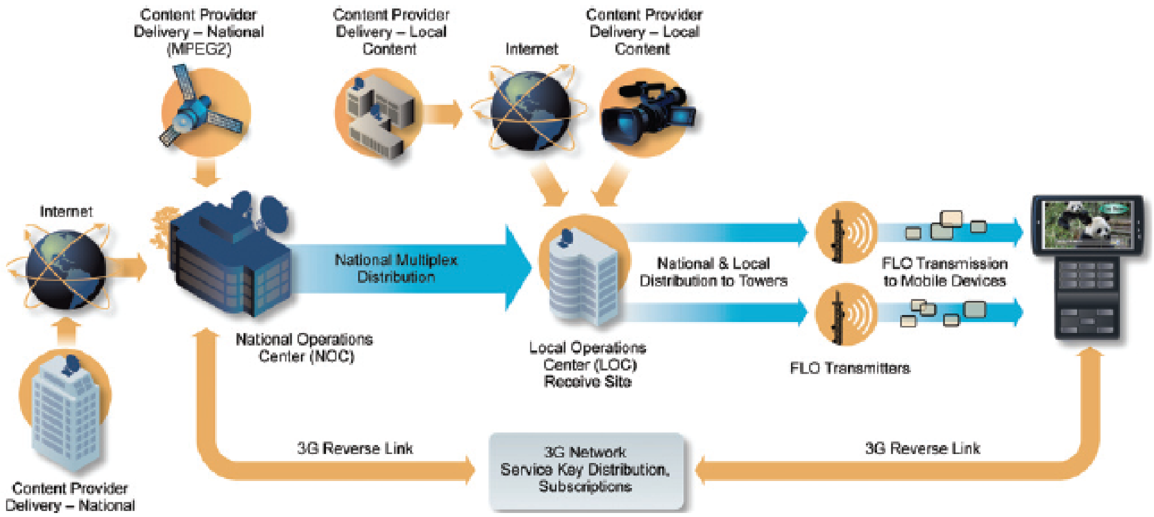

Figure 14: Example of FLO network

Through the clever use of modern wireless engineering, MediaFLO packs an enormous amount of programming into this channel:[73]

> For example, a FLO-based programming lineup that utilizes 30 frames-per-second (fps) QVGA (a Quarter Video Graphics Array or 240x320 pixels) with stereo audio includes 14 real-time streaming video channels of wide-area content (ex: national content) and 5 real-time streaming video channels of local market-specific content. This can be delivered concurrently with 50 nationwide non-real-time channels (consisting of pre-recorded content) and 15 local non-real-time channels, with each channel providing up to 20 minutes of content per day.

A single FLO transmitter, operating at 50 kW effective radiated power, covers an area of 750 square miles. The system is highly robust, as it utilizes OFDM, Reed-Solomon coding, and a purpose-built Medium Access Control protocol, all of which are more advanced and more suitable for mobile TV that the current U. S. standard for DTV that relies on a much less efficient air interface. The downside of the system is limited device support, as each FLO receiver must contain hardware and software support tailored to the service.

5.3.2. Example Video Streaming System: MobiTV

An alternate system built by MobiTV and others utilizes the Mobile/Handheld standard from the Advanced Television Systems Committee (ATSC-M/H). ATSC-M/H piggybacks on frequencies allocated by the FCC to local TV broadcasters. Participating stations simulcast an ATSC-M/H program as a secondary channel carried within their existing 19.39 Mbps 8-VSB DTV stream. ATSC-M/H is less battery power- and bandwidth-efficient than FLO, but the stream is cheap for broadcasters to deploy; MobiTV estimated the cost of adding ATSC-M/H to an existing broadcast system in 2008 at $70,000.[74]

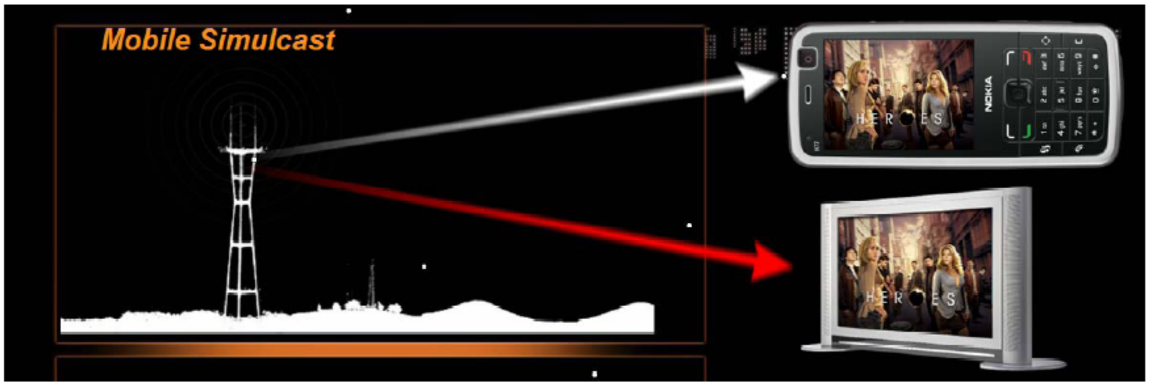

Figure 15: ATSC-M/H Simulcast

A similar system is already in use in Japan, where it's supported by a number of mobile different handsets. A study commissioned by the National Association of Broadcasters estimates as many as 130 million people may be using such systems by 2013 despite their evident technical shortcomings and the burden of standardization.[75]

FLO and ATSC-M/H are both examples of using the bi-directional mobile network to control access to rich media services delivered through a secondary system that performs its one task more efficiently than a general-purpose network can.

FLO and ATSC-M/H are both examples of using the bi-directional mobile network to control access to rich media services delivered through a secondary system that performs its one task more efficiently than a general-purpose network can. They resolve the 3G spectrum crunch by using bandwidth at other frequencies that would otherwise be idle. Regardless of which emerges as the winner, these systems demonstrate that one way of solving the mobile Internet's capacity problem is to use it for the functions it does well, and to bypass it or supplement it for everything else. These approaches are consistent with contemporary mobile device design, which employs multiple radios to increase utility and avoid technology lock-in.

5.4. *Mobile Augmented Reality*

Mobile Augmented Reality (AR) is an exciting new application for handheld devices. The first technology conference devoted to AR will take place in conjunction with the eComm Emerging Communications conference in San Francisco this April.[76]

5.4.1. Example: Layar Platform

Layar is a platform tool that transforms a smartphone into an AR browser; it's a free mobile phone application that displays digital information in real time over pictures take by the smartphone's camera.[77]

The augmented reality experience starts with picture of the user's surrounding which is annotated with information about local features obtained from the Internet. Pointing the camera of a Layer-enabled smartphone at the Abbey Road zebra crossing location made famous by The Beatles opens up a set of AR clues to the 42-location Beatles Discovery Tour. The platform recognizes the images and GPS location, and the application does the rest.

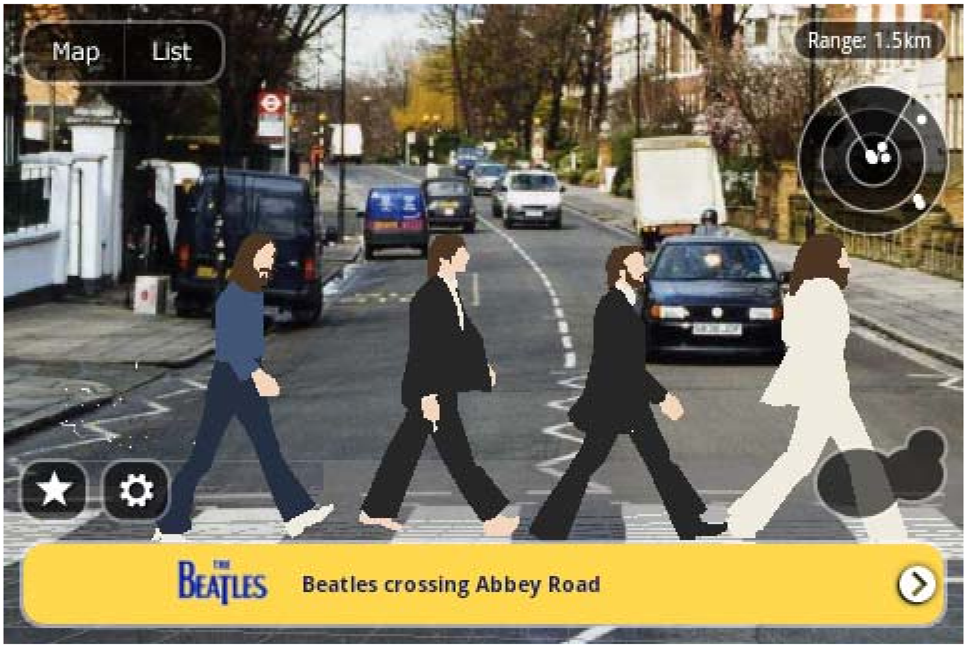

Figure 16: Beatles Discovery Tour starting at Abbey Road.

Virtual Reality systems of the past, which required users to wear cumbersome helmets and goggles, substituted a synthetic experience for a real one, as the Star Trek holodeck did. AR systems supplement reality with insight about local features contributed by others, aiming to make the experience of moving through the world more rich and rewarding. Applications of this sort are enabled by mobile networks, and are literally inconceivable without the base of experience innovators have acquired by creating GPS navigation systems and social networks.

The wide-scale deployment of AR is contingent on the resolution of privacy issues, but it's also dependent on sufficient network capacity to move still images and media clips in both directions between Internet databases and smartphones. These networking requirements are not too challenging, and are essentially consistent with the engineering and policy tradeoffs embedded in mobile networks today. The rise of AR is likely to discourage mobile networks from moving too far from the balance of choices they contain today. Evolved Social Networks face many of the same issues as AR, but without the mobility aspect.

5.5. *Internet of Things*

The Internet of Things is a general category of applications built on machine-to-machine (M2M) communication.[78] If we imagine a world in which very many electronic devices include a network interface and a unique identifier, we can visualize a number of applications in which they communicate with each other. When a dishwasher, smart electric meter, refrigerator, and central heating system can communicate with each other, electricity usage can be scheduled and coordinated to minimize CO_2 production and power bills. Many M2M applications will be low volume and local, with minimal impact on mobile networks and the Internet; others, such as those related to security cameras and baby monitors, will generate large amounts of data that will move through both parts of the Internet, the stationary and the mobile.

A key feature of the M2M applications is the absence of a human decision-maker controlling communications costs; an M2M event occurs when it occurs, and the network deals with it as HAL 9000 dealt with life support in the movie *2001* (somewhat more rationally, we hope.) M2M underscores the importance of applications using communication procedures that adapt to current conditions, with the option of consuming high capacity even under conditions of high load for important events. Reporting a burglary in progress to the local police takes precedence over downloading the latest sitcom, at least from the perspective of the resident of the burgled property.

6. Mobile Internet Policy Issues

Mobile Internet policy revolves around two primary sets of issues, net neutrality and spectrum. In this section, we'll address the primary concerns in both areas.

6.1. *Net Neutrality*

At the State of the Mobile Net conference in Washington last April, discussants in the final panel engaged on the subject "What Policy Framework Will Further Enable Innovation on the Mobile Net?"[79] One viewpoint, urged by public interest advocate Ben Scott and others, calls for the same regulatory framework to be adopted across the mobile and fixed portions the Internet to protect consumers from network operator-induced harm. The basis for this approach is speculative, of course, as we've yet to see an abuse of consumer freedom by a U. S. network operator that wasn't self-corrected in the face of bad publicity or sanctioned by the FCC. The only real net neutrality violation we've yet experienced in the United States was the Madison River case, which was quickly corrected by FCC pressure.

Secondly, even if such abuse were to occur, the notion of an undifferentiated regulatory framework applying to networks ranging from FiOS to the iPhone is hard to justify. Users may perceive mobile and wireline networks as a seamless whole, and may expect transparent and uniform interactions with networks of all types, but networks have technical constraints that will always enable some applications to work better than others most of the time. Providing a uniform personal experience across technologies with widely different constraints is far from straightforward, and writing detailed regulations for unachievable technical goals is counter-productive.

6.1.1. Transparency-based Net Neutrality Lite

That being said, there's no need to return to a technology silos world where the same service offered over different networks has wildly different regulations. It's possible to devise a single technology-independent framework as long as it's sufficiently restrained; the framework can't impose the performance expectations and properties of faster and more robust networks on slower and less robust ones. This approach is sometimes called "net neutrality lite."[80]

We start with the first three of the "Four Freedoms" laid out in former FCC chairman Michael Powell's Internet Policy Statement of 2005:[81]

1. To encourage broadband deployment and preserve and promote the open and interconnected nature of the public Internet, *consumers are entitled to access the lawful Internet content of their choice*.
2. To encourage broadband deployment and preserve and promote the open and interconnected nature of the public Internet, *consumers are entitled to run applications and use services of their choice*, subject to the needs of law enforcement.
3. To encourage broadband deployment and preserve and promote the open and interconnected nature of the public Internet, *consumers are entitled to connect their choice of legal devices that do not harm the network*.

These principles are transformed into restraints on Internet Service Provider conduct rather than consumer entitlements in the Open Internet NPRM:[82]

1. Subject to reasonable network management, a provider of broadband Internet access service may not prevent any of its users from sending or receiving the lawful content of the user's choice over the Internet.
2. Subject to reasonable network management, a provider of broadband Internet access service may not prevent any of its users from running the lawful applications or using the lawful services of the user's choice.
3. Subject to reasonable network management, a provider of broadband Internet access service may not prevent any of its users from connecting to and using on its network the user's choice of lawful devices that do not harm the network.

The NPRM's restatement generally clarifies the intent of the principles, and as such is an improvement. It leaves questions of quality and quantity open, however, and this is bound to lead to conflicts when specific cases are examined.[83]

The NPRM adds two more principles, an anti-discrimination rule and a transparency requirement. We reject the anti-discrimination rule as written:[84]

> "Subject to reasonable network management, a provider of broadband Internet access service must treat lawful content, applications, and services in a nondiscriminatory manner."

It lacks as much as a "reasonableness" qualifier, and ignores the fact the "discrimination," a generally bad thing, is ambiguous in the networking context with "differentiation," a generally good thing. Following the European Commission, we embrace transparency as the cornerstone principle.

6.1.2. Transparency

The principle of transparency assures consumers of the right to full disclosure of all meaningful restrictions on their use of a service, regardless of the network technology in question or the reason for the restriction. Technology changes, and as network technology changes, so do management practices. It's always possible to keep posted management practices up to date with the state of the network, however.

This rule requires interpretation, however. It's possible for network operators to supply users with reams of information about network conditions, most of which is not needed; some practices are trade secret in detail, but not in effect; and some practices experimental and confined to sub-networks while other are system wide.

We recommend the formation of a Technical Advisory Group to work with the FCC on the formulation of crisp and precise transparency and disclosure guidelines.

6.1.3. Content Freedom

Returning to the first three principles, content freedom is generally constructive, but as with any regulation there are limits. The law has a role to play in terms of blocking access to content that fails to meet standards, such as child pornography, sites that traffic in the unlawful sale of licensed material, and sites that propagate malware. As we argued in our report, "Steal These Policies: Strategies for Reducing Digital Piracy," it's generally beneficial for search engines and ISPs to work with voluntary organizations such as Stop Badware to limit access to suspected botnet injectors, and for ISPs to notify users when their computers are infected with malware.[85]

6.1.4. Application Freedom

Applications make themselves known to networks by the patterns of traffic they generate, so "net neutrality heavy" advocates' support of application freedom can easily become a stalking horse for a dysfunctional theory of network management, the "all packets are equal" notion previously cited in reference to Free Press and David Isenberg.[86] This dubious notion is especially harmful to the Mobile Internet because it's disruptive to phone calls, which require low-latency service regardless of the network's current state.

Overall Mobile Internet access efficiency is increased dramatically by the effective management of mobile packets. While net neutrality heavy advocates argue for a command-and-control system of network regulation demanding that all packets be treated indiscriminately, the relevant factor for mobile users is the quality of experience. Myriad management practices and reliability techniques are already practiced from the application level down to the bit level in mobile systems, such that an all-packets-are-equal mandate isn't enforceable in any case.

Management practices should be fully disclosed and placed under user account control as much as possible, but they should be used in any case, as they've always been used in mobile networks.

Packets from a user close to a radio tower are more privileged than those from a user far away, and no regulatory body can change this fact or its many implications. It's also the case that the short-term service impairments common on the mobile network also occur (just less frequently) on wired networks. Because all packet-switched networks are subject to impairment, it's less productive to exempt mobile networks from net neutrality regulations than to write the regulation in such a way that they're cognizant of the real issues in operating packet-switched networks. Mobile networks must be managed more actively than wired ones to operate at peak efficiency, but all networks must be actively managed.

Network operators are in a unique position to harmonize the needs of the users of any given radio tower, each with specific application needs, and available resources. The vision of the edge-managed "Stupid Network" assumes an ever-increasing inventory of bandwidth that can never be exhausted; the reality of mobile networks is an inventory of 410 MHz shared by a population whose appetite for bandwidth doubles every year, obstinate local governments who capriciously withhold permits for new tower sites, limited spectrum, congested backhauls, and a dizzying array of new applications.

In a realistic scenario, harmonizing application needs by raising and lowering priorities, booking bandwidth reservations for isochronous applications, and applying congestion-based pricing measures are only prudent. These practices should be fully disclosed and placed under user account control as much as possible, but they should be used in any case, as they've always been used in mobile networks. Net neutrality heavy would make these practices difficult if not impossible to implement, severely impairing the functionality of the mobile network and unnecessarily reducing the effectiveness of wired networks as well.

This doesn't mean that Mobile Internet users should lack the right to run the applications of their choice or that application providers should be arbitrarily excluded; within the limits of network capacity and service plans and respecting network management systems, there's no sound reason for restricting users' network access on an application-by-application basis. Operators can and do manage on the basis of packet volume by user as well as by traffic patterns, and this is certainly reasonable. The implication of ignoring the application and only considering the volume of traffic it generates is that some applications won't work well all of the time; this is the case on the wired Internet, and it's even more often the case on the more fully utilized mobile Internet.

Effective, application-neutral network management requires mobile network operators to delay and discard specific packets generated by high-volume applications such as peer-to-peer file transfer more frequently than wired network operators do, so the interpretation of "reasonable network management" in the mobile sphere is more expansive than in the wired sphere.

6.1.5. Reasonable Limits on "Device Freedom"

Net neutrality heavy seeks a bright-line regulation banning agreements for exclusive agreements between smartphone manufacturers and network operators, such as the exclusive right granted by Apple to AT&T for distribution of the iPhone in the United States. The net neutrality heavy statement on device freedom is the "Wireless Carterfone" standard advocated by Columbia University Law professor Tim Wu, opposing partnerships between network operators and handset producers.[87]

Net neutrality heavy advocates insist that exclusive marketing agreements conflict with the third principle of the Policy Statement. This right is unfortunately not operative when the user doesn't happen to own the device in question, so the advocates have inverted its intent. We read it as declaring that the consumer should be able to buy whatever device she can find on the open market and bring it to the network, provided the device technology is up to the network's technical standards. Our report on wireless handsets – "Sharing the Risks of Wireless Innovation" – examines this issue in detail.[88]

The device freedom provision is also somewhat troubling on other grounds, due to the fact that it inherits Carterfone notions that were devised to open up a closed, self-managed network to device competition at the edge. A telephone, telephone answering machine, or dial-up modem doesn't produce the same kinds of effects on the PSTN that a device can produce on the end-to-end Internet; in the PSTN case, the network's control logic is wholly internal, and in the Internet's case, it's largely contained in the endpoint. The mobile network invests even more intelligence in the endpoint than the wired Internet does, since the endpoint has to perform a variety of power control, modulation, and coding decisions that go far beyond the capabilities of IP. Before a handset can be cleared for access to the mobile network, the operator has to ensure that it's going to be well behaved under the range of circumstances that affect other users and the network.

> The device freedom notion must be leavened with a great deal of consideration for the responsibility invested in the handset by the wireless network for management and effective operation of the overall network

So the device freedom notion must be leavened with a great deal of consideration for the responsibility invested in the handset by the wireless network for management and effective operation of the overall network.

6.1.6. Managed Services

The NPRM excludes a class of so-called "managed services," which we take to encompass such things as cable TV-like Triple Play packages including voice and video. These are non-controversial exceptions to Open Internet guidelines, and operators suggest that additional managed services are on the horizon providing for better home security, telehealth, distance learning, and telecommuting. It would be beneficial for competition and innovation if a third way of providing remote services existed alongside the classical Internet service and managed services, which we can call "enhanced Internet services." This category of innovation would be enabled by the full implementation of the Internet standards DiffServ and IntServ, which are meant to provide Quality of Service differentiation to applications running across the Internet. We expect that these service enhancements will be charged to the parties who require them, either service providers or end users, on a reasonable and non-discriminatory basis, and that their availability will increase user choice in application-level services. Access to enhanced services on this basis is a constructive form of network unbundling that promotes investment and innovation.

6.1.7. Enabling Innovation

Innovation takes place at different levels in the mobile Internet, each of which may be stimulated by different policy frameworks. The innovation that users see most directly is application innovation. The conventional wisdom of the past held that "open" systems with well-documented application program interfaces (APIs) would give rise to the greatest diversity of applications, and that closed, proprietary systems could only retard innovation. Advocates of this approach to application innovation have been very critical of Apple's management of the iPhone's App Store, most especially the "permission" requirement that Apple must give before applications can be installed on the iPhone. Nevertheless, the approval process is hardly onerous, as we can see from the fact that Apple has approved more applications for the iPhone than have been written for all other smartphones combined. The evidence clearly shows that this model of enabling innovation is a clear success. The success of the iPhone app store suggests that well-managed innovation spaces can produce high levels of innovation despite their supposed lack of "openness" in some dimensions.

The Apple approval process has precedents and counterparts. Microsoft has a testing and approval process for hardware device drivers, the parts of the personal computer system that manage plug-in cards such as graphics processors. Microsoft's process, called Windows Hardware Quality Labs (WHQL) certification, ensures that third party add-ons don't compromise system performance, integrity, or reliability. The WHQL process was once largely optional, but has become essentially mandatory, as Microsoft has raised the bar on security and stability. The Wi-Fi Alliance operates a voluntary certification lab for Wi-Fi hardware devices, and the FCC conducts mandatory certification of Wi-Fi hardware devices.

One the primary reasons that firms such as Apple and Microsoft test and certify is to enable improvements in the software they provide. If third party vendors conform to Apple's user interface guidelines, a single improvement on Apple's part automatically improves each application. Guidelines also ensure a consistent user experience, reducing the learning curve for each application.

Apple and Microsoft have a unique vantage point with respect to the innovative new uses of the systems they provide, but in principle it's not much different from the vantage point of any platform provider. Facebook has a unique awareness of the breadth and nature of Facebook plug-ins, and network operators are aware of the nature of network traffic and therefore able to optimize in ways that no one else can.

The global viewpoint enjoyed by operators also makes it possible for each firm and agency in a permission-granting capacity to abuse its power. This raises the inevitable question of checks and balances, one of the central tussles in regulation theory. The European solution is instructive as it emphasizes consumer choice and consumer information above detailed government mandates that can easily become quite cumbersome to define and enforce. The European view is that the ability of consumers to change wireless providers is the strongest pressure against arbitrary practices.

6.1.8. Specific Example: EU Telecoms Package

One specific way of resolving the network regulation dilemma recently emerged in Europe. The EU Telecoms Package imposes a disclosure requirement on network operators. The approach is broadly permissive, and does not yet impose the minimum service level on network operators some advocates had sought. European guidelines resemble the FCC first three freedoms plus the transparency rule:[89]

> For the European Commission, the open architecture of the Internet is of key importance for the Information Society. The Commission in particular considers that the following "net freedoms" should be general guidelines for regulators and policy makers: right for users to access and distribute (lawful) content, to run applications and connect devices of their choice.
>
> The Commission therefore proposes, in the EU Telecoms reform, a transparency mechanism concerning possible restrictions on consumers' choice of lawful content and applications so that consumers can make an informed choice of services and reap the full benefits of technological developments. In practice, consumers will get clear and timely information from their service providers about any restrictions that the providers place on their access to or use of Internet or mobile content and applications. This will allow them to pick and switch to the operator which best suits their needs. Where consumers have no alternative, service providers should not be allowed to block or restrict such access.

Despite a great deal of support for an anti-discrimination rule the European Commission stopped short of imposing one. However, it did reserve the right of national regulators to impose such a rule should circumstances warrant. The European Commission also imposed a rule requiring network operators to implement a one-day switching procedure for customers changing mobile providers. Hence, disclosure and switching perform the anti-discrimination function in Europe, without a specific anti-discrimination rule and its many problems of definition and application. If there is to be an anti-discrimination rule, effective enforcement requires the FCC to organize a Technical Advisory Group of technical experts and network operators as well as public interest advocates to examine and interpret specific management practices.

6.1.9. Services-Oriented Framework

The addition of the disclosure and switching rules to the European regulatory corpus is consistent with past European practice that regards telecom services as uniform commercial practices regardless of the technology that implements them, a departure from the traditional American practice of treating each technology as a silo guided by unique regulations.[90] While each technology has its own capabilities and limitations, the dynamics of disclosure, basic rights, and rapid switching are uniform across all of them. Consumer services should live up to claims made for them, whether implemented across advanced networks or not.

Networks are platforms for the services provided by applications providers and content publishers, and these applications are themselves platforms for additional services. The services-oriented approach can thus apply equally to a dominant service provider in any part of the Internet ecosystem, mobile or fixed, hardware- or software-oriented, content- or communications-driven. The services model does not discriminate for or against specific industries, and it does not pick winners and losers. Historically, the Internet ecosystem has enabled a wide variety of interactions and business models, and it's clearly in the public interest for it to continue to enable such innovations in the future. In the European view, a minimum of ex-ante rules, the disinfectant of full disclosure and easy switching is the best hope for preserving Internet diversity. It's also the case that the "regulation of services not of technologies" model is more constructive than a "consumer rights" model bounded only by permanent limitations of specific technologies. Most of the practical limitations that occur in wireless networks occur in wired networks, only less frequently. Therefore, the management freedom needed by wireless operators is needed by wired network operators as well.

6.2. *Spectrum Policy*

Spectrum policy encompasses the terms under which network operators and users are permitted to use various parts of the electromagnetic spectrum. These terms of use cover power levels in each frequency band, the contours of the electromagnetic waveforms that transmitters are allowed to generate, and access procedures such as sensing energy from high-priority users or checking a database before transmitting. These rules generally permit two forms of access: an unconditional use license bounded only by power limits, and license-free uses for devices that pass a certification procedure that ensures compliance with guidelines.

Spectrum policy is in part a tussle between the licensed and license-exempt regimes, but an additional element has emerged recently in the form of an additional competition between licensed users of broadcast TV spectrum and licensed users of the mobile network. In the relevant frequencies for mobile networking (between 400 MHz and 4 GHz), significant allocations have been made for TV broadcasting and White Spaces in addition to the approximately 410 MHz auctioned to mobile network operators.

6.2.1. The Value of Open Spectrum

Regulatory models for license-exempt spectrum use are called Open Spectrum models. Open Spectrum models are beneficial for spectrum uses that propagate or cover limited areas. Wi-Fi, for example, is required to conform to such a low transmit power limit that signals transmitted by a Wi-Fi Access Point operating at the common 2.4 GHz frequency inside a house dissipate within a block, and those transmitted at the 5.8 GHz frequency called used by the 802.11a variety of Wi-Fi seldom penetrate more than one wall. Signals of this kind have so little potential to cause interference that the trouble of issuing and enforcing licenses is not worthwhile. Millions of people around the world use Wi-Fi as an alternative to cabled Ethernet for gaining access to home, office, and campus networks providing Internet service and other utility functions such as printer sharing.

Wi-Fi is a raging success story, but its cousin Ultra-Wideband (UWB) has so far not been a success. Unlike Wi-Fi, which uses “junk bands” shared by baby monitors, garage door openers, cordless phones, and microwave ovens, UWB uses enormous swathes of spectrum, at least 500 MHz, ranging from 3.1 to 10.6 GHz, but it emits so little power so infrequently than its signal is effectively invisible to licensed uses. Unfortunately, the power limit set by the FCC is so low that it’s hard build a practical UWB device. This is an experiment in secondary use of spectrum, from which we’ve learned the pitfalls of constructing regulations too conservatively.

White Spaces is another experiment in Open Spectrum which is too early in the process to deliver definitive findings. Most advocates believe its power limits were also set too conservatively by the FCC, which may lead to a fate similar to UWB unless corrected. White Spaces takes the novel approach of including a database that devices must consult to learn about authorized frequencies, and all White Spaces use is subject to an experimental license that must be reauthorized every 24 hours.

6.2.2. The Limits of Open Spectrum

From the engineering perspective, the most significant difference between licensed and license exempt wireless networks is coordination. Wireless signals don’t propagate uniformly; hence any wireless network has to deal with the effects of coverage overlap. Licensed systems are engineered for centralized management across multiple cells, but license-exempt systems following open standards such as IEEE 802.11 presume that each cell is separately managed. IEEE 802.11 systems try to avoid interference with neighboring systems by seeking less crowded frequencies; with only three or four channels to choose from in the 2.4 GHz band, uncrowded channels don't always exist. The 802.11 standard unfortunately lacks a general means of coordinating across access points administered by different users.

The fact that standards for inter-access point coordination don't exist doesn't imply that they can't be developed; 802.11 systems employ a spectrum sharing etiquette packet-by-packet, and inter-access point coordination is simply spectrum sharing at a slightly larger scale. Ultra-Wideband networks conforming to the WiMedia Alliance's MAC protocol standard employ a spectrum sharing system that allows overlapping networks to reserve airtime from a common pool and to assign it to specific sessions that require QoS guarantees.[91] It's a very effective system for small-scale networks with minimal overlaps (UWB has a 30 foot effective signal range), but is not designed to scale to the neighborhood level and above.

The issue that ultimately sabotages open spectrum systems in high usage scenarios isn't the technical challenge of allocating spectrum according to a pre-determined algorithm; it's the deeper challenge of aligning the incentives of users. Licensed systems ultimately manage on a price-per-use model, where operators and subscribers are engaged in continual negotiation. Paid networking models contain built-in incentives for users to moderate their bandwidth demands and for operators to implement fair sharing. When this constraint is relaxed, as it is in free networking models, another constraint must arise in order to encourage users to share fairly.

This constraint typically arises from the resources accessed by users of free networks; most commonly, these are shared gateways to the wireline Internet infrastructure or paid content. Wireline gateways don't always perform this function efficiently (they have so much capacity per dollar it doesn't matter,) so the burden of incentivizing fair sharing falls on content. When content is free – as it us in the case of pirated songs and movies, genuinely user-generated content and public interest content – there's no effective incentive for individual users to moderate their bandwidth demands other than good character and public spiritedness.

Economic research on commons suggests that users will develop informal systems of cooperation to maintain the long-term health of shared resources such as grazing land and water, but such systems depend on cross-communication and reputation, two commodities that are rare in networks where anonymous users barter pirated content.[92] Ultimately, the reason that technical standards have not developed meaningful systems for cross-network resource sharing is that such algorithms must be specified by policy makers, not simply by engineers. Crossing the cultural rift between the tech policy and engineering communities to devise and enforce such systems of spectrum sharing is a difficult task, but it must be done if open spectrum models are to flourish on a large scale.

6.2.3. What Licensed Spectrum Does Well

Licensed spectrum models handle two problems extremely well: coordination among overlapping users, and the replacement of obsolete systems with state-of-the-art equipment that benefits from recent and ongoing advances in technology. The value of these factors was discussed in the section on Wireless Infrastructure.

6.2.4. Spectrum Inventory and Review

The International Telecommunications Union (ITU) estimates that the United States will require 800 MHz of mobile spectrum by 2010; other estimates are similar. So far, U.S. regulators have only licensed approximately 410 MHz, somewhat less than the current European allocation. European regulators are making more spectrum available to their network operators: Germany is auctioning an additional 340 MHz, the UK 262 MHz, and France is in the middle of a series of 3G auctions that will be followed by 4G auctions freeing up an additional 90 MHz in all.[93]

The need for mobile spectrum comes from the widespread adoption of smartphones capable of running advanced applications. Cisco estimates the volume of mobile data will double every year through 2014, increasing 39 times between 2009 and 2014, for a compound annual growth rate (CAGR) of 108 percent. By their estimate, total traffic will reach 3.6 exabytes per month by 2014.[94]

Regardless of the status of spectrum licenses, wireless engineering continues to improve spectral efficiency. Morgan Stanley predicts LTE will achieve 16 bits/Hertz/tower in the field.[95] While this estimate is lavishly optimistic (Merrill Lynch puts the figure under 3 bits/Hertz/tower, disregarding sectorization), there is a clear trend toward greater system spectral efficiency.[96]

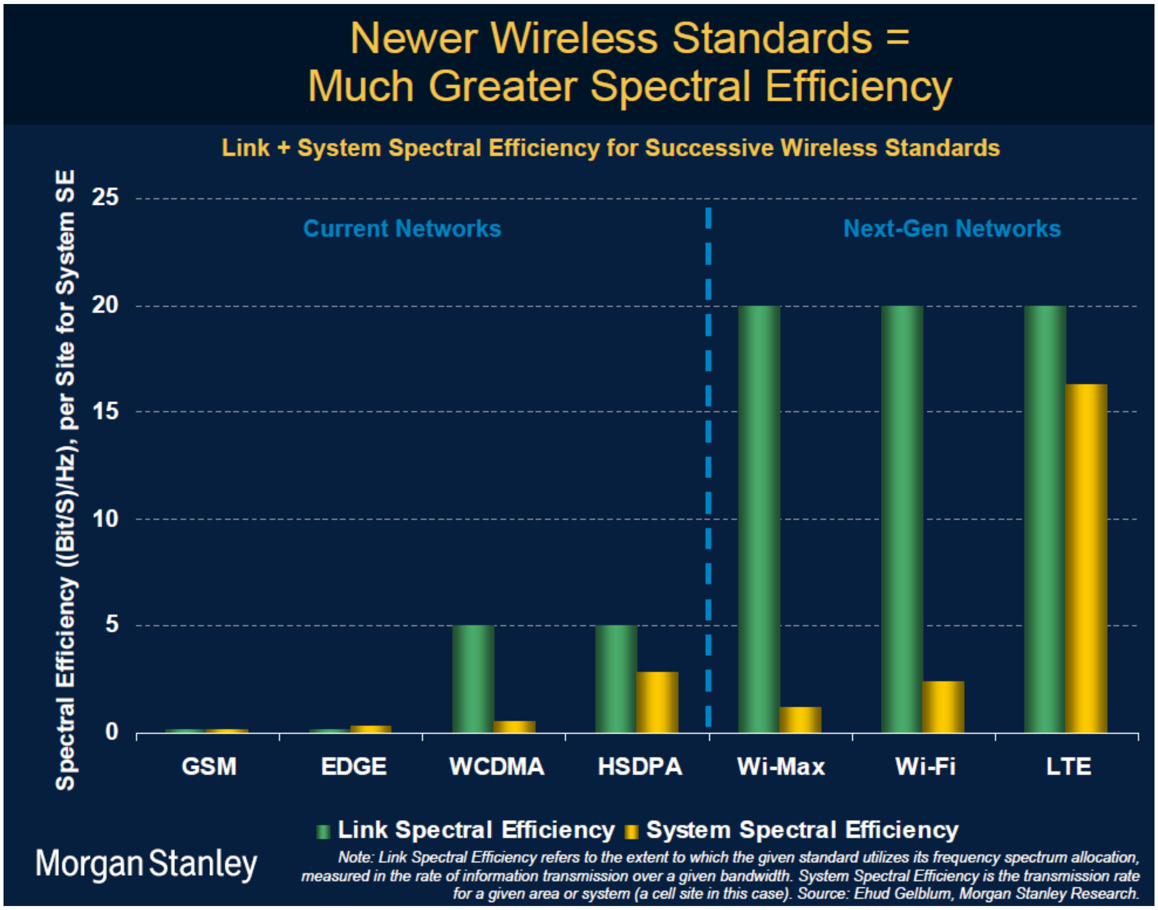

Figure 17: Licensed LTE is more efficient than license-exempt Wi-Fi.

System efficiency increases most rapidly in licensed bands due to the effective exemption they have from contention overhead, as we've explained. However, increased spectral efficiency in existing allocations isn't sufficient to meet the needs of emerging applications and widespread adoption of smartphones. Over the long term, wireless spectral efficiency doubles every 30 months, but user demand for bandwidth roughly doubles every 11 months. The only way out of this dilemma is to make more spectrum available for mobile use, in either licensed or license-exempt forms, or in both.

In addition to the 410 MHz of choice licensed spectrum, the U. S. currently makes some 650 MHz of license exempt spectrum available below 6 GHz, including the White Spaces bands, the 2.4 GHz "Wi-Fi 802.11b/g" band, the 5.8 GHz "Wi-Fi 802.11a" band and the 3.6 GHz "Wi-Fi 802.11y" band.

The traditional Wi-Fi allocations alone amount to 475 MHz (75 at 2.4, and 400 at 5.8), and an additional 50 MHz of Wi-Fi spectrum exists in the 3.6 GHz 802.11y band for limited use. White Space rules potentially cover all of the 360 MHz allocated for Over-the-Air TV not assigned to broadcasters or reserved as guard bands next to active broadcast channels. The 650 MHz of license-exempt spectrum below 6 GHz exceeds the 410 MHz licensed to mobile network operators, but not all frequencies are equal.

The primary value of the lower end frequencies such as White Spaces is generally thought to arise in stationary point-to-multipoint and mobile applications in sparsely populated rural settings. In densely populated urban settings, the unpredictable propagation characteristics at 700 MHz make dense packing of towers challenging, but this issue may ultimately be ameliorated by advanced engineering techniques. Given the heavy bias evident in current allocations toward license-exempt uses and the system inefficiency of license-exempt technologies in high-density areas, it's difficult to argue that even more spectrum should be made available under such terms.

The argument for additional license-exempt spectrum is based on the expectation that the progeny of spread spectrum radio (SDMA, beam forming, Software-Defined Radios, and Cognitive Radios,) will some day enable mobile users to exploit unused frequencies wherever they happen to be. Assuming this prediction is correct, we still need to know when it will come to pass. Will these technologies, which have been touted since the spread spectrum experiments of the 1960s, come on-line quickly enough to satisfy the capacity needs of mobile users, or will unrealistic predictions of their immediate utility simply delay the release of licensed spectrum? Additionally, is there any reason to believe that license holders will not move to adopt any and all techniques that increase the efficiency of the networks they operate?

Farber and Faulhaber argue that the hoped-for methods of opportunistic spectrum use will not be practical for quite some time:[97]

> Technologies such as cognitive radio that are potential users of shared spectrum are very far from actual deployment; there is no need to act precipitously to establish rules for technologies whose eventual implementations may be very different from today's views.

Existing license-exempt bands should be sufficient to prove the utility of next-generation radio technologies, and once they cross the threshold to practicality, regulators can revisit the question of licenses. This dynamic is no different from historical practice, and it should become more common as new radio technologies emerge and existing practices become obsolete. It probably becomes wise for regulators and license holders to view licenses as long-term rental agreements rather than permanent grants in any case.

6.2.5. Reassigning Spectrum

The United States' current DTV standard – ATSC – was obsolete before it was deployed, having been chosen fifteen years ago before OFDM and MIMO were well understood. Its inefficiency has prevented us from reaping a digital dividend as large as the one that can be anticipated in Europe and the rest of the world. This is an unfortunate fact that can't be immediately corrected because consumers have made substantial investments in ATSC and can't very well be expected to undergo another DTV transition in the near future. There are several means of mitigating this regulatory error, however. The most straightforward is to encourage TV broadcasters to share spectrum with each other and return unused or lightly used spectrum to the FCC for auction.

Instead of two TV channels separated by a white space broadcasting separate programming streams to a diminishing audience, broadcasters should be encouraged to share common streams for essential programming. They should then return excess spectrum to an auction pool, from which it can be reassigned to mobile networks, mobile TV using efficient means such as MediaFLO, or other bidders representing growing markets. In return for this sacrifice, it's reasonable that they should be compensated with a share of auction proceeds.

Systems such as that proposed by CTIA for replacing high-power transmissions with networks of lower-powered transmissions should be examined for technical merit and potentially adopted. Finally, policymakers should consider the utility of replacing over-the-air TV with subsidized cable and satellite TV services, especially in low population density areas. Ultimately, the entitlement to free over-the-air television must be reconsidered, perhaps over the very long term.

A parallel problem exists with the use of spectrum by government agencies. One way to encourage efficient use of government spectrum is to create a government spectrum czar's office with the power to review, reassign, and transfer government uses to more modern means.

These two initiatives are examples of an overarching spectrum modernization program.

6.2.6. Future Proofing the Spectrum

What sort of regulatory framework best promotes the development of innovative new technologies with the potential of improving society, extending and improving life, and generally making the world a better place? This question underlies the debate about Internet regulation in Washington and elsewhere. Ultimately, we fall back on philosophy in constructing an answer because we can't predict the future. The FCC's historical approach, until the advent of auctions, was to identify the best and highest use for each frequency band all them allocate them accordingly. This approach led to allocations that didn't fare well with advances in technology, extending creaky systems such as AM radio well beyond their meaningful life. Permanent allocations deny the fundamental property of technology, improvement over time.

High-end mobile devices, including both small and large form factors, employ multiple radios today: 3G, Wi-Fi, Bluetooth, FLO, and GPS are each radio frequency receivers, and all but GPS and FLO are transmitters as well.

Mobile platforms with multiple radios are the first widely adopted devices to be immune from problems caused by spectrum re-allocation. They tend to have a short useful life (on average consumers replace handsets every 18 months,) and they're usable in multiple modes. The Apple iPod touch is essentially an iPhone without a 3G radio, and its Wi-Fi capability permits it to be used on any 3G network accessible by Wi-Fi. The pocket-sized, battery-powered Novatel Mi-Fi device attaches Wi-Fi-capable devices to Verizon's and Sprint's CDMA networks, and in an alternate form attaches to European GSM/HSPA networks.

Figure 18: Novatel Mi-Fi Device

Voice capability isn't as always as good through the Mi-Fi as it is with the iPhone's native 3G interface, but other functions are comparable.

The first step toward future proofing the spectrum is to create devices that can accommodate reallocation of spectrum either as consequence of regulatory reassignment or because of repurposing by spectrum auction winners. Commercial interests can obviously move faster and more effectively in improving spectrum efficiency than government can, if for no other reason than the fact that commerce is able to employ trial-and-error techniques unsuited to the government scenario that includes a broad pool of stakeholders with uncertain commitment to particular technical approaches.[98] One widely publicized approach to spectrum flexibility is the software-defined radio (SDR), a device that can become one of the many elements of a multiple radio platform as soon as SDR technology makes necessary advances in power consumption and antenna efficiency. SDR will most likely become an invaluable means of ensuring efficient spectrum use one day, but it's simply not there yet.

Now that we have mobile devices capable of operating at multiple frequencies, using optimal modulation, coding, and MAC protocols on each, we can make changes more rapidly in the way that spectrum is assigned and operated than we could in the past, actually decommissioning obsolete uses very rapidly. Future spectrum allocation models can be more software-oriented and digital than the models of the past. Smart mobile devices make software-oriented spectrum policy practical.

Each network technology permits a range of innovation within the scope of the network's capability. For obvious reasons, HDTV streaming did not emerge on narrowband networks and interactive gaming is not employed across high-latency satellite links. There is always a tension between emerging applications and network capabilities. Application innovators always demand more, and network operators decide how best to improve networks along the multiple dimensions of capacity, delay, and price. Innovation is messy and unpredictable, often taking place on multiple frontiers simultaneously. If we ask innovators how networks should improve, they'll each ask for features that improve their own application. It's therefore up to operators to mediate conflicting desires.

7. Policy Recommendations

The technology developments and regulatory debates around the mobile Internet are indicators of its social value; movements don't form and coordinated international dialogs don't emerge around issues that no one cares about. Basic questions of technology and regulation may not ever be fully resolved, but essential principles and action steps are clear.

The general approach we recommend is for the government to facilitate the Mobile Internet by removing impediments to further build-out and adoption. Speculative fears have played too large a role in the Internet regulation debates of the last decade, and it's more productive to shift the emphasis toward the government's role in facilitating progress.

First, it would be a mistake to impose the "net neutrality heavy" guidelines on either wired ISP networks or mobile networks. Rather than enacting overly prescriptive regulations against experimenting with new transport service and business models, the FCC should rely primarily on transparency and disclosure to protect consumers from speculative harms, maintain active oversight of provider practices, and reserve direct intervention for instances of clearly harmful conduct. Second, policymakers should embark on a program of spectrum modernization and expansion to ensure that mobile services can continue to grow. A special focus should be placed on the transfer of licenses from inefficient DTV use to the general pool of spectrum available for auction. Spectrum modernization should also be employed to replace inefficient federal, state and local government uses and release unneeded spectrum to an auction pool. Finally, regulations should encourage technical solutions to be developed and deployed that enable consumers to obtain the best possible service for the best prices. Doctrinaire net neutrality heavy formulas simply don't accomplish that end within the constraints of mobile networks.

7.1. *Stick with Light-touch Regulation*

Heavy-handed regulation is ultimately bad for investment, deployment, and adoption of wireline networks, and potentially fatal to mobile networks. A key way to ensure that networks serve the public interest is through market mechanisms based on meaningful competition. The United States currently enjoys among the most competitive intermodal wireline broadband and even stronger wireless competition, with four national wireless networks, as well as a number of regional networks and Mobile Virtual Network Operators (MVNOs) such as Virgin Mobile. Fixed wireless networks, such as the Clearwire system, are reasonable substitutes for wireline, and the two satellite networks are in the process of upgrading capacity significantly. Competition can be made more effective by ensuring there are minimal delays in switching between mobile providers.

7.2. *Enact a Sensible Transparency Rule*

While the European transparency rule is too new to be able to be judged a success, it represents a promising direction for which there is broad consensus. There is still disagreement regarding the specific nature of required disclosure, which is understandable given the complexity of network systems and the gap between consumer awareness and technology. Just as a well-functioning democracy requires an informed citizenry, a well-functioning network ecosystem requires its well-informed and honest critics. The challenge for a transparency rule is to disclose the things that must be disclosed in order for users to gauge the experience they'll have on any given part of the Internet ecosystem in terms the average person can understand, while making additional information available to the audience of technologists and policy analysts. Certain details of practice represent trade secrets and need not be disclosed; the means by which a particular user-visible effect is produced are less important than the effect itself. One approach that recommends itself is the co-regulatory approach championed by Marsden, in which stakeholders convene with the regulator to draft specific guidelines.[99] Toward that end, we encourage stakeholders to form a working group to advise the FCC on the particulars of disclosure.

7.3. *Legitimize Enhanced Transport Services*

There is widespread agreement among filers in the FCC's Open Internet NPRM that differentiated services for differentiated fees are legitimate in their own right, and not simply as an adjunct to network management. Similar services have a long history on the Internet, where they are known as Content Delivery Networks, Overlay Networks, and Transit Networks. Extending the logic of "pay more to get more" down to the level of individual user accounts has an obvious appeal. These practices have proved worthwhile for content publishers and resellers such as Netflix and Skype, so it stands to reason that they would be beneficial for future competitors in the market for video streaming and telephony. If ISPs who operate "eyeball networks" are permitted to compete with CDNs and Overlays, new entrants can expect lower prices and more competition, and end users should expect a wider array of options.

7.4. *Define Reasonable Network Management*

The transparency rule, and its specific implementation, provides insight into the boundaries of reasonable network management practices. While the use of the term “reasonable” without definition is impossibly vague, anchoring management practices to service disclosure resolves a great deal of the mystery. We know that a practice is reasonable if it does what the operator says it does, conforms to standards devised by responsible bodies such as IEEE 802, IETF, and the ITU, and doesn’t violate basic user freedoms. We know that it’s unreasonable if it fails to accomplish its stated purposes and restricts user rights in the process. Beyond these general guidelines, a Technical Advisory Group must work with the FCC to develop additional clarity regarding management boundaries.

7.5. *Preserve Engineering and Operations Freedom*

The primary emphasis of the Open Internet NPRM’s framework of rules is on the preservation of users’ freedom to experience the Internet as they see fit, without arbitrary limitations. A key way to preserve this freedom is to address the dynamics of technical freedom that make it possible. Users experience the Internet as they do now because engineers, network operators, and application innovators have been free to improve networks, network technology, and user experience.

Toward that end, the NPRM should make it clear nothing in the FCC’s approach denies the freedom to invent, develop, and adopt new networking technologies, business models, and practices that have the potential to enhance the Internet’s power, efficiency, vitality, or effectiveness.

The FCC should consider adding two additional principles to its list: Engineering Freedom and Operations Freedom. The telephones that worked on the PSTN in the first year of the Carterfone regime still work 35 years later. If the cell phones we use today are still usable on the mobile network 35 years from now (or even ten years from now), that should be regarded as a failure of innovation. The Mobile Internet is driven by an ethic of continual improvement and this principle more than any other must remain in the forefront. Thus, we propose two additional rules for the Open Internet NPRM:

- No part of this regulation shall be construed as limiting the freedom of network engineering to devise, develop, and deploy technologies to enhance the Internet or to improve user experience.
- No part of this regulation shall be construed as limiting the freedom of Internet Service Providers, other network operators, or other service providers to devise new financial or business models that better align user incentives with those of network operators or application-based service providers without limiting user choice.

These rules make it clear that innovation is the engine that best ensures the Internet’s continued public value.

7.6. *Review Existing Spectrum Licenses*

The FCC needs to complete its inventory of the licenses it has issued over the years, and implement a system that eliminates or reduces ambiguity about licenses going forward. If it's true that the FCC has somehow lost track of some licenses, as some have suggested, this error should be corrected. It's simply not acceptable for the national regulator of wireless networks to lose track of issued licenses. The Kerry spectrum map is a step in the right direction.

7.7. *Eliminate Redundant and Archaic Licenses*

Once the license inventory is complete, it will be possible to examine licenses to determine which are unused, which are redundant, and which can be combined with others to free up spectrum for auction or other kinds of assignment. Part of this process will entail reassigning some occasional uses to the control of other agencies, license holders, or custodians of other kinds. Rarely used public safety applications can be combined with consumer services, for example, by allowing public safety uses to take precedence in times of emergency. The general principle that should hold in the process of review is modernization, replacing archaic analog applications with more spectrum-efficient digital ones. No single approach to spectrum management exceeds all others in terms of general utility, but there should be a bias in favor of spectrum custodians in either the public or the private sector with vested interests in efficient use. Sufficient spectrum exists, in principle, to meet projected user requirements for mobile networking. There is not sufficient spectrum that we can afford to waste large swathes on speculative projects of uncertain utility, however. A reasonable approach is embodied in the White Spaces order, where all licenses are experimental ones renewable day-by-day. Proven applications can be rewarded under this system with license of longer duration.

7.8. *Protect Spectrum Subleasing*

Secondary markets for licensed spectrum enabled by resale and subleasing have proved useful in the U. S., where dozens of Mobile Virtual Network Operators (MVNOs) lease capacity from license holders and roaming agreements permit licensees to share capacity. These kinds of secondary markets are also useful in the microwave backhaul and point-to-point space where a given license holder can adjust microwave paths with relays and dogleg arrangements to accommodate most effective use.

7.9. *Cautiously Enable Secondary Uses*

One area of controversy concerns such secondary uses as wireless underlay and overlays on licensed spectrum. Advocates insist that such uses are non-interfering with properly restricted, and license holders are skeptical. The reality is that the nature of the interference caused by overlay networks such as Ultra-Wideband depends on the nature of the incumbent service. Ultra-Wideband interferes, in some installations, with highly sensitive applications such as radio astronomy, but this fact is known and the Ultra-Wideband waveform is adjusted accordingly. When the details of the incumbent service are known, in terms of coding, modulation, and framing protocols, overlay and underlay services can be engineered for cooperation without interference. Nevertheless, when details of the primary service change, interference may arise anew. For this reason, all secondary uses should be required to back off and even shut down completely until they can be certified as non-interfering with the primary license holder. The principle use of secondary services should be in areas where the primary user is not active; this is the logic behind the Dynamic Frequency Selection (DFS) system in IEEE 802.11a Wi-Fi. This system requires Wi-Fi systems to look for the use of radar on certain channels, and to refrain from using channels where radar is found.

In all cases, the burden falls on the secondary user to avoid causing interference with the primary user. Systems of enforcement for this principle need to be incorporated into all secondary use regulations; the White Spaces database has this capability.

7.10. *Allow the Experiment to Continue*

The Internet as we know it today is the fruit of a 35-year experiment. In the beginning, it was the prototypical science project, albeit one with government support shepherded by a highly skilled and dedicated band of researchers, champions, and developers out to prove that a new vision of networking was not only practical but superior to the old one.

The mobile data network has a completely different creation story, originating in a commercial context and targeted toward adding an important new feature to the existing network without fundamentally altering its nature.

Each of these networks has a story, a set of champions, and a vision. Each has been transformative in its own way, giving rise to its own industry, and liberating some vital element of human society along the way. It's not surprising that the convergence of these networks should occasion debate and conflict, some of it intense and heated.

The way forward requires some give and take. It's not enough to impose the Internet's operational traditions on the mobile network, because the Internet's operational community has chosen not to adopt the Internet standards most relevant to mobile networking: RSVP, IntServ, and Mobile IP. It's not enough for mobile operators to demand that Internet users abandon open access to the web at reasonable speeds in favor of a constrained system of locked-down portals and proxies. Each culture has things to learn from the other.

The way forward is a careful, diligent, step-by-step process beginning with reviews of historical rules and precedents and ending in the creation of a new framework that will enable the next generation of networking to flourish. The evidence of an emerging consensus among responsible parties in the U. S. and Europe suggests it's well underway.

Acknowledgements: The author thanks Dale Hatfield, Charles Jackson, and Peter Ecclesine for invaluable assistance with the technical content of this report; all factual errors are the responsibility of the Information Technology and Innovation Foundation. The author thanks Daniel Castro, Emiko Guthe, and Priscilla Jang for editorial and production assistance.

---

[1] Christopher T. Marsden, *Net Neutrality: Towards a Co-regulatory Solution* (London: Bloomsbury Academic, 2010), http://www.bloomsburyacademic.com/pdf files/NetNeutrality.pdf.

[2] A previous report provides a full description of the development of the Internet's technical architecture and the role of end-to-end arguments: Richard Bennett, *Designed for Change: End-to-End Arguments, Internet Innovation, and the Net Neutrality Debate* (ITIF, September 2009), http://www.itif.org/index.php?id=294.

[3] A fourth network, the U. S. Military's AUTODIN II was later included in the design space, but never deployed.

[4] The adoption of the datagram-oriented and layered CYCLADES architecture followed the rejection of two earlier ideas. The first attempt at Internet design placed a gateway between each pair of networks, an approach that was rejected because it wouldn't scale. The second attempt combined TCP and IP into a single protocol, an approach that was rejected because it would have impaired real-time applications such as voice.

[5] L Pouzin, *The Cyclades Computer Network: Towards Layered Network Architectures* (Amsterdam; New York: North-Holland, 1982). Bob Kahn was employed by BBN before coming to ARPA to lead the "Internetting" effort. He and colleagues at BBN collaborated with Pouzin on the CYCLADES design.

[6] This process is described in a previous work by the author: Bennett, *Designed for Change*.

[7] David Weinberger, *Small Pieces Loosely Joined: A Unified Theory of the Web* (New York: Basic Books, 2002).

[8] The forum was the First ACM symposium on Operating System Principles held in Gatlinburg, Tennessee in October 1967. The seminal work on modular design was Edsger W. Dijkstra, "The structure of the "THE"-multiprogramming system," in *Proceedings of the ACM symposium on Operating System Principles - SOSP '67* (presented at the ACM symposium, Gatlinburg, 1967), 10.1-10.6, http://portal.acm.org/citation.cfm?doid=800001.811672. The seminal work on ARPANET was Lawrence G. Roberts, "Multiple computer networks and intercomputer communication," in *Proceedings of the ACM symposium on Operating System Principles - SOSP '67* (presented at the ACM symposium, Gatlinburg, 1967), 3.1-3.6, http://portal.acm.org/citation.cfm?doid=800001.811680.

[9] J. Postel, "RFC 795 - Service mappings" (Network Working Group, September 1981), http://tools.ietf.org/html/rfc795. Two of the networks, ARPANET and AUTODIN II, had high-priority service options, and other two had low delay routing options. AUTODIN II, conceived as a replacement for the AUTODIN military message-passing network, was never deployed; see: Jared Hall, "Autodin," *Website of Jared Hall*, October 14, 2005, http://www.jproc.ca/crypto/autodin.html.

[10] Bennett, *Designed for Change*.

[11] Internet congestion control corrected a condition that caused throughput to drop to less than one percent of design capacity, so by contrast 30 percent efficiency is wonderful.

[12] Unsecured SNMP information is sometimes helpful in stemming DDoS attacks. See: Joseph Menn, *Fatal System Error: The Hunt for the New Crime Lords who are Bringing Down the Internet*, 1st ed. (New York: PublicAffairs, 2010): *[Barrett Lyon] had just finished a weekend surf session [when] it hit him: at least some [zombie hosts] had to be using a basic piece of networking software called the Simple Network Management Protocol*.

[13] *TR-069: CPE WAN Management Protocol v1.1*, Technical Report (The Broadband Forum, December 2007), http://www.broadband-forum.org/technical/download/TR-069_Amendment-2.pdf.

[14] See H-W. Braun, "RFC 1104 - Models of policy based routing" (Network Working Group, June 1989), http://www.faqs.org/rfcs/rfc1104.html., especially Section 6. Policy based dynamic allocation of network

resources (e.g., bandwidth, buffers, etc.): *The policies involved in making dynamic bandwidth allocation in a largely packet switching environment possible are still in the development phase. Even the technical implications of infrastructure reconfiguration in result of events happening on a higher level still requires additional research… The technical problem of sharing network resource policy information could be solved by a making a "network resource policy information" database available to all administrators of networks and Administrative Domains. However, the political problems involved in creating a network resource policy with impact on multiple Administrative Domains does still require additional study.*

[15] D. Meyer, L. Zhang, and K. Fall, "RFC 4984 - Report from the IAB Workshop on Routing and Addressing" (Network Working Group, September 2007), http://www.ietf.org/rfc/rfc4984.txt.

[16] "[The Internet] was a common market of innovation, protected by an architecture that forbade discrimination." Lawrence Lessig, *The future of ideas: the fate of the commons in a connected world*, 1st ed. (New York: Random House, 2001), p. 85.

[17] P2P protocols randomize port numbers in an attempt to evade detection; since no other protocol randomizes, this measure is dead giveaway.

[18] The Merriam-Webster dictionary lists "google" as a transitive verb meaning "to use the Google search engine to obtain information about (as a person) on the World Wide Web." See: http://www.merriam-webster.com/dictionary/GOOGLE

[19] The web works by transferring a series of files from web server to web browser; the most common Internet innovations are web applications.

[20] A Princeton University census of files available on Mainline DHT, a trackerless BitTorrent index, indicates that 99% are copyright infringing. See: Ed Felten, "Census of Files Available via BitTorrent," blog, *Freedom to Tinker*, http://www.freedom-to-tinker.com/blog/felten/census-files-available-bittorrent.

[21] Andy King, "The Average Web Page - Preliminary Results," *Optimization Week*, http://www.optimizationweek.com/reviews/average-web-page/.

[22] BitTorrent, Inc. is working on a new peer-to-peer protocol called uTP which is claimed to depart from historical peer-to-peer behavior, but some have doubts as to its effectiveness. See: George Ou, "BitTorrent would rather be selfish than friendly," Blog, *Digital Society*, http://www.digitalsociety.org/2010/01/bittorrent-would-rather-be-selfish-than-friendly/.

[23] Paul J. Feldman, "Court Challenges FCC in Early Network Neutrality Test," *CommLawBlog*, January 8, 2010, http://www.commlawblog.com/2010/01/articles/internet/court-challenges-fcc-in-early-network-neutrality-test/.

[24] Iljitsch van Beijnum, "Playing chicken: ISP depeering a high-school lovers' quarrel," news blog, *Ars Technica*, March 21, 2008, http://arstechnica.com/old/content/2008/03/isps-disconnect-from-each-other-in-high-stakes-chicken-game.ars.

[25] Roger O. Crockett, "At SBC, It's All About "Scale and Scope"," news magazine, *Business Week*, November 7, 2003, http://www.businessweek.com/@@n34h*IUQu7KtOwgA/magazine/content/05_45/b3958092.htm.

[26] J. Scott Holladay and Inimai M. Chettiar, *Free to Invest: The Economic Benefits of Preserving Net Neutrality*, Institute for Policy Integrity (New York: New York University School of Law, January 2010), http://www.policyintegrity.org/documents/Free_to_Invest.pdf. This analysis is flawed on many grounds, of course.

[27] Marsden, *Net Neutrality: Towards a Co-regulatory Solution*.p. 91.

[28] E. Rosen, A. Viswanathan, and R. Callon, "RFC 3031 - Multiprotocol Label Switching Architecture," Internet RFC, January 2001, http://tools.ietf.org/rfc/rfc3031.txt.

[29] S. Blake et al., "RFC 2475 - An Architecture for Differentiated Services," Internet RFC, December 1998, http://tools.ietf.org/rfc/rfc2475.txt.

[30] R. Braden, D. Clark, and S. Shenker, "RFC 1633 - Integrated Services in the Internet Architecture: an Overview," Internet RFC, June 1994, http://tools.ietf.org/rfc/rfc1633.txt.

[31] H. Schulzrinne et al., "RFC 1889 - RTP: A Transport Protocol for Real-Time Applications," Internet RFC, January 1996, http://www.ietf.org/rfc/rfc1889.txt.

[32] H Schulzrinne, A Rao, and R Lanphier, "RFC 2326 - Real Time Streaming Protocol (RTSP)," Internet RFC, April 1998, http://www.ietf.org/rfc/rfc2326.txt.

[33] R Braden et al., "RFC 2205 - Resource ReSerVation Protocol (RSVP)," September 1997, http://www.faqs.org/rfcs/rfc2205.txt.
[34] Within the network operator community, rumor has it that Global Crossing, Level 3, TeliaSonera, and two German networks negotiate such agreements, but these rumors can't be confirmed with public information. It is clear that a number of ISPs offer QoS guarantees directly to their residential and business customers, which tends to imply broader agreement among ISPs.
[35] The word "transitive" is used in routing discussions in the sense of the "transitive relation" taught in seventh-grade algebra: "Whenever an element *a* is related to an element *b*, and *b* is in turn related to an element *c*, then *a* is also related to *c*." - "Transitive relation," in *Wikipedia*, http://en.wikipedia.org/wiki/Transitive_relation.
[36] The ideas in this section are mainly taken from Mark Handley, "Why the Internet only just works," *BT Technology Journal* 24, no. 3 (July 2006): 119-129. http://www.cs.ucl.ac.uk/staff/M.Handley/papers/only-just-works.pdf. Handley is Professor of Networked Systems at UCL and author of 22 Internet RFCs including SIP, the protocol that performs telephony signaling.
[37] Ibid.
[38] Meyer, L. Zhang, and Fall, "RFC 4984."
[39] Sources: Takehiro Nakamura, "Proposal for Candidate Radio Interface Technologies for IMT□Advanced Based on LTE Release 10 and Beyond (LTE□Advanced)" (3GPP, October 15, 2009), http://www.3gpp.org/IMG/pdf/2009_10_3gpp_IMT.pdf, Glen Campbell, "Mobile data: Traffic jam ahead?" (Bank of America - Merrill Lynch, February 2, 2010), others.
[40] Cellular telephony inventor Martin Cooper discovered this eponymous law, which he says has held since Marconi's time.
[41] Exactly how it does telephony has not yet been decided, and may not be for some time. While these details are developed, LTE networks will "fall back" to earlier standards to complete phone calls.
[42] The energy from the local transmitter is so strong from the receiver's location that the receiver would not be able to detect interference from a far source.
[43] A pair of enhancements to the 802.11n Medium Access Control protocol called "frame aggregation" increase efficiency for some streaming applications.
[44] For this reason, Wi-Fi itself has an optional mode called "802.11e Scheduled Access" that functions more like a mobile network.
[45] Tim Wu, "Network Neutrality, Broadband Discrimination," *SSRN Electronic Journal* (2003), http://www.ssrn.com/abstract=388863.
[46] Steve Deering, "Watching the Waist of the Protocol Hourglass" (presented at the IETF 51 Plenary, London, August 2001), http://www.ietf.org/proceedings/51/slides/plenary-1/sld001.htm.
[47] R. Braden, Clark, and Shenker, "RFC 1633 - Integrated Services in the Internet Architecture: an Overview."
[48] R Braden et al., "RFC 2205 - Resource ReSerVation Protocol (RSVP)."
[49] AT&T, Orange, Telefonica, TeliaSonera, Verizon, Vodafone, Alcatel-Lucent, Ericsson, Nokia Siemens Networks, Nokia, Samsung Electronics Co. Ltd., and Sony Ericsson, "Global Telecom Companies Announce A Standards Based Solution For Voice And SMS Services Over LTE," Press Release, November 4, 2009, http://news.vzw.com/news/2009/11/pr2009-11-03a.html.
[50] "Deutsche Telekom Announced First Voice Calls over LTE with VoLGA," *Cellular News*, December 9, 2009, http://www.cellular-news.com/story/41004.php.
[51] The One Voice coalition just gained the support of GSMA. See: Phil Goldstein, "GSMA sides with One Voice via VoLTE effort," news magazine, *Fierce Wireless*, February 15, 2010, http://www.fiercewireless.com/story/gsma-bolsters-support-one-voice-lte-approach/2010-02-15.
[52] Diagram from: Bala Thekkedath, "IMS-Enabled Access Networks: Enabling network-independent access to personalized services," White Paper (UT Starcom), http://www.utstar.com/Document_Library/0639.pdf.
[53] Jim Courtney, "Skype Business Model Revealed at eBay Analyst Event," news blog, *Voice on the Web*, March 12, 2009, http://voiceontheweb.biz/2009/03/skype-business-model-revealed-at-ebay-analyst-event/. Skype and BitTorrent are both privately held firms that don't make financial data public, but most analysts believe Skype is profitable and BitTorrent isn't.

---

[54] These priorities are described as Network Control, Voice, Video, Controlled Load, Excellent Effort, Best Effort, and Background. See: LAN MAN Standards Committee of the IEEE Computer Society, "IEEE Std 802.1D - 2004" (The Institute of Electrical and Electronics Engineers, Inc., June 9, 2004), http://standards.ieee.org/getieee802/download/802.1D-2004.pdf, p. 241.
[55] These priorities are described as Voice, Video, Best Effort, and Background. See: LAN MAN Standards Committee of the IEEE Computer Society, "IEEE Std 802.11e - 2005" (The Institute of Electrical and Electronics Engineers, Inc., November 11, 2005), http://standards.ieee.org/getieee802/download/802.11e-2005.pdf, p. 1.
[56] Janna Anderson and Lee Rainie, *The Future of the Internet IV* (Washington, DC: Pew Research Center's Internet & American Life Project, February 19, 2010), http://pewinternet.org/Reports/2010/Future-of-the-Internet-IV/Part-4Architecture.aspx?r=1.
[57] "Nation's First White Spaces Network Deployed in Rural Virginia," company blog, *The Spectrum Bridge*, November 4, 2009, http://spectrumbridge.blogspot.com/2009/11/nations-first-white-spaces-network.html.
[58] "Free-space path loss," in *Wikipedia*, http://en.wikipedia.org/wiki/Free-space_path_loss.
[59] See: "Concealed Cell Sites," *Larson Utility Camouflage*, http://www.utilitycamo.com/sites.html.
[60] Phil Goldstein, "FCC approves tower siting 'shot clock' rules," *FierceWireless*, November 18, 2009, http://www.fiercewireless.com/story/fcc-approves-tower-siting-shot-clock-rules/2009-11-18?utm_medium=rss&utm_source=rss&cmp-id=OTC-RSS-FW0.
[61] Dale Hatfield, personal communication.
[62] Google's citywide Wi-Fi project in Mountain View, California uses GigaBeam systems for backhaul. GigaBeam is bankrupt.
[63] "Ceragon Networks Ltd.," corporate web site, February 11, 2010, http://www.ceragon.com/default.asp.
[64] Gigabeam Corporation, "One Velocity Application Spotlight," http://gigabeam.com/LinkClick.aspx?fileticket=z484C63qBbs%3d&tabid=86.
[65] Virgin Mobile offers a 3G USB modem for $99.99 and pay-as-you-go broadband access for as little as $10. See: "Mobile Broadband Service," *Virgin Mobile*, http://www.virginmobileusa.com/mobile-broadband.
[66] David S. Isenberg, "The Dawn of the Stupid Network," *ACM Networker* 2.1, February/March 1998: 24-31, http://www.isen.com/papers/Dawnstupid.html.
[67] Applications that use UDP regulate their own bandwidth demands to align them with the network's capacity at any given instant; TCP applications are regulated by network signals about peak capacity sent by dropping packets.
[68] Bernhard Warner, "On test: the 3 Skype phone," newspaper, *Times Online*, November 14, 2007, http://technology.timesonline.co.uk/tol/news/tech_and_web/personal_tech/article2871385.ece.
[69] "3 Skypephone S2 is absolutely pathetic! :-@," *Skype Community*, October 27, 2009, http://forum.skype.com/index.php?showtopic=455981.
[70] Warner, "On test: the 3 Skype phone."
[71] Mark Sullivan, "AT&T Wireless CEO Hints at 'Managing' iPhone Data Usage," magazine, *PC World*, http://www.pcworld.com/article/173320/atandt_wireless_ceo_hints_at_managing_iphone_data_usage.html.
[72] "Slingbox app approved by AT&T, just in time for iPad launch," on-line newspaper, *The Christian Science Monitor*, February 4, 2010, http://www.csmonitor.com/Innovation/Horizons/2010/0204/Slingbox-app-approved-by-AT-T-just-in-time-for-iPad-launch.
[73] "FLO Technology Overview" (Qualcomm, Inc., 2009), http://www.mediaflo.com/news/pdf/tech_overview.pdf, p. 4.
[74] "ATSC-M/H: The Promise of Free to Air Mobile Simulcast" (MobiTV, 2008), http://www.mobitv.com/technology/whitepapers/ATSC.PDF, p. 1.
[75] Richard V. Ducey and Mark Fratrik, *Study of the Impact of Multiple Systems for Mobile/Handheld Digital Television* (Washington, DC: BIA Financial Network, 2008).
[76] Lee S. Dryburgh, "2010-2020 Will be the Decade of Mobile Augmented Reality: Day Three Now Full!," blog, *Emerging Communications*, February 11, 2010, http://blog.ecomm.ec/2010/02/2010-2020-decade-of-augmented-reality.html.
[77] "Augmented Reality Browser," company web site, *Layar*, http://layar.com/.

[78] See: "Internet of Things," in *Wikipedia*, http://en.wikipedia.org/wiki/Internet_of_Things.
[79] "State of the Mobile Net Conference 2009" (Conference, Washington, DC, April 23, 2009), http://www.netcaucus.org/conference/2009/sotmn/agenda.shtml. video of final panel available at: http://www.youtube.com/watch?v=FE_ECfzP57I.
[80] Marsden, *Net Neutrality: Towards a Co-regulatory Solution*.
[81] Federal Communications Commission, "Policy Statement," September 23, 2005, http://hraunfoss.fcc.gov/edocs_public/attachmatch/FCC-05-151A1.pdf. The fourth freedom relates to consumer choice and is therefore of a different nature.
[82] Federal Communications Commission, "Notice of Proposed Rulemaking, GN Docket No. 09-191; WC Docket No. 07-52," October 22, 2009, http://hraunfoss.fcc.gov/edocs_public/attachmatch/FCC-09-93A1.pdf.
[83] Does the right to "run" an application such as Skype mean that my call quality will always be good, or does it mean I must accept the fact that I may find that my call inaudible? If the instantaneous demand for network bandwidth exceeds the supply and differentiation is forbidden, this outcome is inevitable.
[84] Federal Communications Commission, "Notice of Proposed Rulemaking, GN Docket No. 09-191; WC Docket No. 07-52."
[85] Daniel Castro, Richard Bennett, and Scott Andes, *Steal These Policies: Strategies for Reducing Digital Piracy* (Washington, DC: Information Technology and Innovation Foundation, December 15, 2009), http://www.itif.org/index.php?id=324. "StopBadware," *Badware Busters*, http://stopbadware.org/; Elinor Mills, "Comcast pop-ups alert customers to PC infections," news magazine, *CNet News*, October 9, 2009, http://news.cnet.com/8301-27080_3-10370996-245.html?tag=newsEditorsPicksArea.0.
[86] M. Chris Riley and Robb Topolski, "The Hidden Harms of Application Bias" (Free Press/New America Foundation, November 2009), http://www.freepress.net/files/The_Hidden_Harms_of_Application_Bias.pdf. Isenberg, "The Dawn of the Stupid Network."
[87] Tim Wu, "Wireless Carterfone," *International Journal of Communication* 1, Columbia Public Law Research Paper No. 07-154 (2007): 389, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=962027.
[88] Richard Bennett, *Sharing the Risks of Wireless Innovation* (Washington, DC: Information Technology and Innovation Foundation, October 19, 2009), http://www.itif.org/index.php?id=303.
[89] Europa, "7 Very Concrete Improvements for European Consumers," press release, http://europa.eu/rapid/pressReleasesAction.do?reference=MEMO/08/552&format=HTML&aged=0&language=EN&guiLanguage=en.
[90] Bennett, *Designed for Change*.
[91] Wi-Fi's most commonly used mode, EDCF, allocates bandwidth through packet-by-packet contention; WiMedia UWB allocates over a longer timeline, stream-by-stream. See: "WiMedia MAC Specification 1.5" (WiMedia Alliance, Inc., December 1, 2009), http://www.wimedia.org/imwp/download.asp?ContentID=16552.
[92] Elinor Ostrom, *Governing the Commons: The Evolution of Institutions for Collective Action* (Cambridge; New York: Cambridge University Press, 1990). documents the development of systems of coordination for non-networked resources but not of communication networks themselves.
[93] For Germany and UK see: Mary Meeker et. al., "The Mobile Internet Report: Key Themes" (Morgan Stanley, December 15, 2010), http://www.morganstanley.com/institutional/techresearch/pdfs/Mobile_Internet_Report_Key_Themes_Final.pdf. For France see: Michelle Donegan, "French 4G Spectrum Update," news blog, *Light Reading Mobile*, January 13, 2009, http://www.lightreading.com/blog.asp?blog_sectionid=414&doc_id=170295.
[94] "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2009-2014" (Cisco, February 9, 2010), http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-520862.pdf.
[95] Meeker et. al., "The Mobile Internet Report: Key Themes."
[96] Campbell, "Mobile data: Traffic jam ahead?"
[97] Gerald R. Faulhaber and David J. Farber, "Mandated Spectrum Sharing: A Return to "Command and Control"" (Comments filed with FCC), http://fjallfoss.fcc.gov/ecfs/document/view?id=7020246460.

[98] Gerald R. Faulhaber and David J. Farber, "Innovation in the Wireless Ecosystem: A Customer-Centric Framework," *International Journal of Communication* 4 (2010): 73-112.
[99] Marsden, *Net Neutrality: Towards a Co-regulatory Solution*.